# DOCUMENT RESUME

ED 156 718 95 TM 007 285

| | |
|---|---|
| AUTHOR | Harris, Chester W.; And Others |
| TITLE | Achievement Test Items--Methods of Study. CSE Monograph Series in Evaluation, 6. |
| INSTITUTION | California Univ., Los Angeles. Center for the Study of Evaluation. |
| SPONS AGENCY | National Inst. of Education (DHEW), Washington, D.C. |
| PUB DATE | 77 |
| NOTE | 157p. |
| AVAILABLE FROM | Center for the Study of Evaluation, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California 90024 ($4.50) |
| EDRS PRICE | MF-$0.83 HC-$8.69 Plus Postage. |
| DESCRIPTORS | Academic Achievement; *Achievement Tests; Annotated Bibliographies; *Bibliographies; *Item Analysis; Item Sampling; *Mathematical Models; Matrices; Standard Error of Measurement; *Statistical Analysis; *Test Reliability; Tests of Significance; Test Validity |

ABSTRACT

The implications of a mathematical model of test scores are explored where the data are limited to a random sample of items without replacement from an indefinitely large population or item domain in which items are scored either zero or one. The purpose is to obtain an unbiased estimate of a student's proportion of items correct in the item universe. The study of item responses plays no role in the test development process, since item difficulty will vary with instruction, and consideration of item-test correlation would bias the item sampling procedure. Three major emphases are: (1) the association or equivalence between pairs of items; (2) the sensitivity of items to instruction, which may also be a function of instructional effectiveness; and (3) the stability of an item after instruction. Item scores are assumed to be a function of a constant plus a value that characterizes an individual, and the error of measurement is analyzed in two terms. The development of the theory is concerned with data summarized meaningfully in two-by-two or fourfold tables such as the data for a Phi coefficient. There is an annotated list of over 100 selected journal articles plus a list of additional bibliographical references. (CTM)

***********************************************************************
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
***********************************************************************

ED156718

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CSE
MONOGRAPH
SERIES
IN
EVALUATION

6

# ACHIEVEMENT TEST ITEMS— METHODS OF STUDY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mark Young

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

CENTER FOR THE STUDY OF EVALUATION
UNIVERSITY OF CALIFORNIA • LOS ANGELES

TM007 285

# CSE MONOGRAPH SERIES IN EVALUATION

**SERIES EDITOR**

**Eva L. Baker**

Center for the Study of Evaluation
UCLA Graduate School of Education
University of California, Los Angeles
Los Angeles, California 90024

# ACHIEVEMENT TEST ITEMS—
# METHODS OF STUDY

Chester W. Harris,
Andrea Pastorok Pearlman,
and
Rand R. Wilcox

Center for the Study of Evaluation
UCLA Graduate School of Education
University of California, Los Angeles, 1977

## CSE MONOGRAPH SERIES IN EVALUATION

| NUMBER | | |
|---|---|---|
| 1. | Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case from the MINNEMAST Project<br>Wells Hively, Graham Maxwell, George Rabehl, Donald Sension, and Stephen Lundin | $3.50 |
| 2. | National Priorities for Elementary Education<br>Ralph Hoepfner, Paul A. Bradley, and William J. Doherty | $3.50 |
| 3. | Problems in Criterion-Referenced Measurement<br>Chester W. Harris, Marvin C. Alkin, and W. James Popham (Editors) | $3.50 |
| 4. | Evaluation and Decision Making: The Title VII Experience<br>Marvin C. Alkin, Jacqueline Kosecoff, Carol Fitz-Gibbon, and Richard Seligman | $3.50 |
| 5. | Evaluation Study of the California State Preschool Program<br>Ralph Hoepfner and Arlene Fink | $3.50 |
| 6. | Achievement Test Items—Methods of Study<br>Chester W. Harris, Andrea Pastorok Pearlman, and Rand R. Wilcox | $4.50 |

This Project was supported in whole or in part by the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

# TABLE OF CONTENTS

Chapter

Preface . . . . . . . . . . . . . . . . . . . vii

I Introduction . . . . . . . . . . . . . . . . . 1

*Chester W. Harris*
*Andrea Pastorok Pearlman*
*Rand R. Wilcox*

II Data Models . . . . . . . . . . . . . . . . . 6

*Chester W. Harris*

III Conventional Significance Tests and Indices of Agreement or Association . . . . . . . . . . . . 17

*Chester W. Harris*
*Andrea Pastorok Pearlman*

IV New Methods for Studying Stability . . . . . . . . . 45

*Rand R. Wilcox*

V New Methods for Studying Equivalence . . . . . . . . 66

*Rand R. Wilcox*

VI Abstracts of Selected Journal Articles . . . . . . . . 77

*Andrea Pastorok Pearlman*

Bibliography . . . . . . . . . . . . . . . . . . 127

Additional References . . . . . . . . . . . . . . . . 130

# PREFACE

CSE's research and development efforts are directed toward improving the practice of evaluation in educational settings. Our major focus is on the preparation of products for both academic and professional consumers of evaluation.

Since its inception, one of CSE's major efforts has been the exploration of the field of educational measurement. Over the last decade, the level of support for research and development in the field of measurement has been relatively low and has generated only a modest number of findings of impact. Because our program was one of few to receive programmatic support, the history of our efforts in this field is illuminating. At CSE, landmark development of the Institutional Objectives Exchange (Popham, Baker, Skager, 1968), the System for Objectives-Based Assessment in Reading (Skager, 1971), and the continued work in research on bilingual assessment models (Cornejo, 1977) provide instances where recognition of the field needs for improved measurement procedures generally outstripped the methodological base for the developments. Contemporaneous to these development projects, our recognition of the degree of public expenditures in the area of testing led us to the preparation and application of criteria for the evaluation of commercially available tests. Tests covering basic and higher order intellectual skills as well as the affective domain were evaluated for preschool, elementary, and secondary students (Hoepfner, 1972). This kind of work is continuing with our evaluations of available criterion-referenced tests (Walker, 1977). The findings from both of these test evaluation projects strongly suggest a need for an integrated theory of achievement testings. Until such a theory is developed, systematic improvement in the quality of achievement tests is unlikely and our basis for judging progress toward this improvement will be inadequate.

Throughout these years, CSE has conducted research into problems associated with criterion-referenced and norm-referenced measurement, using both conceptual and empirical methods. (See, among others: Klein & Kosecoff, 1974; Baker, 1971).

While our studies have been augmented by those conducted by other members of the national R&D network (at Pittsburgh Learning Research and Development Center, Wisconsin Research and Development Center for Cognitive Learning, and the Southwest Regional Educational Laboragory, for example) aggregating these experiences shows that the level of understanding of tests, their design, their interpretation, and their utilities for classroom and policy decision making are at best rudimentary.

A review of research conducted in the field of measurement is disheartening. While measurement theory has relatively strong psychometric bases it has an almost complete lack of substantive or content underpinnings. Further, the meager substantive theory that is linked to quantitative concerns draws mainly from psychological theory, primarily from the area of personality. We do not think that this traditional psychometric theory is appropriate for achievement tests. A theory of tests content with quantitative models that correspond to the theory is needed for achievement tests.

Although some of the issues that need to be addressed by a theory of achievement tests have been considered they have not been attacked in a programmatic fashion because it was not regarded in the national interest to do so; information needs and training requirements for evaluation personnel were so great that development and dissemination of such relatively primitive notions about measurements as objectives were regarded as quantum leaps for practitioners and was therefore supported. However, there is still the need for long-term attention to measurement theory. This monograph represents a step toward a unified theory of achievement testing. Approaching this problem from a quantitative perspective, the authors are led to conclusions simultaneously emerging from concerns with instruction.

Because of CSE's general focus in the area of achievement, one of our goals has been the development of an integrated theory of achievement testing. In general, research on a theory of achievement testing might be conceived of as having at least four major components: (1) basic problems—comparative, conceptual, and empirical analyses of alternative measurement approaches; (2) research on the design of tests, including the creation of non-trivial rules governing their valid specification for various purposes; (3) development of models for analysis of test adequacy through the use of technical indicators; (4) modeling and experimentation on the interpretation of test results in various environments.

CSE is addressing each of these four areas. We expect that our work will proceed in a configurational rather than linear pattern, with certain lines of inquiry considered in parallel rather than in sequence. It is not within CSE's resources to solve the problem of achievement test theory single-handedly. But through the setting of a nationally disseminated research agenda for scholars and practitioners, an agenda that is shared through conferences and publications such as this, we hope to stimulate the measurement community to address long-ignored but vital issues in the areas of measurement.

We invite your comments.

Eva L. Baker
Director, Center for the Study of
Evaluation

## Chapter I

# INTRODUCTION

The purpose of these papers is to describe a point of view concerning the measurement of achievement and to explore certain consequences of this point of view. This introductory paper attempts to define the context within which the nature and purposes of the measurement of achievement will be considered; this will be accomplished by examining the relationships between instruction and measurement that set limits on the practices appropriate for the construction of achievement measures. We see two types of studies of achievement measures. One type consists of judgmental analyses of the definitions entering the achievement measures, the procedures used in developing the measures, the judged structure of the pool of items or parts used to make up the measures, and the like. These studies are made by presumably informed or expert judges and are made with reference to the instructional program with which the achievement measures are intended to be used. The second type of study examines responses of appropriately defined categories of students to the achievement measures under certain planned conditions and uses methods of data summary, analysis, and interpretation that are specifically designed to answer questions about the functioning of what we shall call the teaching-testing complex. For both these types of studies, sampling procedures may be employed as a practical necessity and, if so, they require that methods of estimation and the development of satisfactory estimators also be considered as relevant problems for study within the two types.

The point of view we develop here depends not only on the intimate linkage of instructional purposes and procedures with measurement procedures, but also on the conception of the relevant data and how it arises. Thus, in the second paper, we turn our attention to data models and attempt to identify models that are consistent with our notion of the role of achievement measurement in connection with instruction and that, at the same time, make minimal assumptions about the structure of the data. This analysis of data models is made at the level of the item and it sets the stage for the consideration of ways of studying student responses to items that will give useful information about the teaching-testing complex. The remaining papers address this question of useful response information in this context and how it should be gathered, summarized, and extrapolated. In these remaining papers we review conventional tests of hypotheses and estimation procedures, many of which are well known. However, our requirements often lead to the rejection, in a particular situation, of an estimator or a test that might otherwise be recommended. These remaining papers also give new results for some estimation procedures that are not well-known and seem not to have been utilized previously in the study of achievement test item responses. For these less well-known estimators, we also examine some of their technical characteristics.

*A point of view.* Our concern is with achievement testing that is linked specifically to an instructional program, and we make no comments about other situations. Furthermore, we do not consider all possible types of achievement tests. These papers consider only achievement tests that can be made up of distinct items and that have as their "natural" metric the proportion of items handled successfully. Thus, we have nothing to say about the single item test, which might be represented by an individual's science project or by an individual's research paper. Instead, we assume that a large (indefinitely large) number of items is or can be made available, and that it is the proportion of these items that a student can handle successfully which is the parameter to be estimated for that student. We also have nothing to say about achievement tests that have a different "natural" metric, such as a test of typing speed which probably should be scored as a rate rather than as a proportion correct. These exclusions do not leave us with an empty set, however, because the achievement requirement in basic skills and knowledges can be conceptualized as sets of distinct items (item domains) for which the proportion correct metric gives a meaningful observation about a student. We recognize that not all instructional programs may at present be describable in sufficient detail to permit a definition of an achievement requirement with regard to an item domain; at the same time, we hope that progress toward such specification can be made in more and more areas of educational achievement.

The question of the ways in which the measurement of achievement is or may be linked to the instructional program deserves more extensive consideration than we undertake here. We merely assume the objectives of the instructional program exist and that it is possible to explicate the objectives so as to yield the specifications of what constitutes the evidence of achievement. These specifications should describe (1) the types of tasks which the student is expected to perform as a result of the instruction, (2) the types of materials or content (including "ideas") that these tasks should involve, and (3) the types of situations in which the behavior is expected to be elicited. We refer to work of Hively and his associates (1973) as an illustration of how a universe of task-plus-content items can be generated from a curriculum design. We note that their item-generation procedure can define an *implicit* universe of items without necessarily producing every one of the items at a given time. Our requirement will be that a universe of items exists either implicitly or actually in such a form that it will be feasible to draw random or stratified random samples from this universe. The Hively item-generation procedure satisfies this requirement.

Our notion of achievement testing determines "how an individual performs at present in a universe of situations that the test situation is claimed to represent" (*Standards for Educational and Psychological Tests and Manuals*, 1966, p.12), with the proviso that this universe of situations (item domain) is directly related to the instructional program in the sense previously mentioned. Given the universe of situations, it then is possible to use a random sample of these items as an achievement test for a single student, and it also is possible to use independent random samples of items for repeated testing of the same student. Shoemaker (1975) argues for a position much like ours and discusses the advantages for individual and group assessment that can accrue from it.

An achievement test constructed on this item sampling principle yields a proportion correct score which is an unbiased (and maximum likelihood) point estimate of the proportion of the items in the domain which the student can handle adequately. It is important to recognize that this is a generalization based on sampling principles. As such, the estimate will be biased whenever the selection of items to make up the test is not random but is based on item analysis procedures which abandon or eliminate items having certain statistical characteristics, such as low internal consistency characteristics or low sensitivity to instruction. We are aware that others, like Messick (1975) for example, are willing to bias this estimate; our position is that a random selection of items from the total item domain rather than one based on item analysis should be used to construct an achievement test.

The logic of our argument is that: A first and critical step in the construction of an achievement test is the specification of the objectives of the instruction and the related specification of the tasks, content, and occasions that define the behaviors that the instructional program is designed to promote. This analysis and specification yields an item universe or item domain, the performance on which would be the measure of achievement for any student at that point in time. This measure, if available, would be the universe or criterion score. The validity of this measure, expressed as a proportion of the item universe which the student handles adequately, is completely dependent upon the adequacy of the analysis of instructional objectives and the specification or generation of the item universe. The concept is closest to the concept of content validity, but it also involves a concept of task and occasion validity. Because the criterion or universe score generally cannot be made available for a student, since it is not feasible for him to respond to every item in the domain, we must adopt an estimation procedure. An obvious procedure is the random sampling procedure we have mentioned, which uses the proportion of "correct" responses to a sample of items as the unbiased point estimate of the individual's parameter—his universe proportion correct. An important feature of this estimation procedure is that it can be carried out for one individual student without access to similar data for other students.

*The role of item response data.* Assuming that our position is correct, the study of item responses—as in reliability estimation and item analysis—plays no role in the test development process. It may, however, play a role in the study of the teaching-testing complex. We turn to this consideration shortly.

We first wish to make the point that for achievement tests linked to an instructional program, the responses of students to the achievement test items are a function of both the instruction and the character of the items. There is a confounding here that cannot be escaped and certainly should not be assumed away. In general, we expect that instruction, which is a variable that can be manipulated, will affect performance on any item set, providing the instruction comes at any appropriate stage for the students; the effects may be many and they may differ at different stages. This means that item difficulties and universe scores can be altered and that no assumption about the distribution of universe scores can be a very comfortable one for long. It also means that conventional relations between items as variables may differ at different stages, and thus, that estimates of reliability may fluctuate in relation to instruction. If so, item response data cannot be properly interpreted in the absence of information about the instructional history of the respondees. (A corollary, of course, is that achievement scores for students cannot be properly interpreted in the absence of information about the item domain.) To be informative, studies of achievement item response data, then, should be made only for samples of carefully defined populations of students, with instructional history a critical variable in defining populations.

It is our position that evidence about the characteristics and/or functioning of items in an appropriate item domain gathered from one or more such populations can provide useful information about the teaching-testing complex. Our view is that it will be useful to characterize the item domain at various points in the instructional process. For example, an obvious expectation is that as instruction proceeds to its conclusion, there will be an increase in the average, over students, of the easiness of the items in the domain. A type of study which we describe later under the label of "sensitivity to instruction"—without ignoring the fact that both item sensitivity and instructional effectiveness are reflected in the data—provides evidence about this expectation for any selected item and for any aggregate of items. Another obvious expectation is that with practice and familiarity the item response tends to become stable—i.e., it is learned. This expectation can be tested and the stability of response estimated for any selected item or for any aggregate of items at any stage of the instruction. We also discuss a third type of study of item response data which considers pairs of items. The expectation is that, as instruction progresses, a larger proportion of the students will respond correctly to any given pair of items that belong to the domain; this in part reflects the increasing easiness of the individual item but it also reflects an association between items. We use the term "equivalence" in discussing this third type of study. These three types of studies can be respectively interpreted as a validity study, a specific reliability study, and a generic reliability study. However, our methods, which are discussed in succeeding chapters, are not necessarily those conventionally associated with these terms.

---

Chapter II

# DATA MODELS

---

**CHESTER W. HARRIS**

We now wish to discuss data models that may be employed in the measurement of achievement in the context of an instructional program. We assume that an item* universe, either implicit or explicit, is available, and that this item universe defines or conceptualizes at least a segment of the student achievement that the instructional program is designed to foster. We also assume that it will be feasible to select, at random, a sample of these items to employ as a "test" for a given individual. We begin with a concern for the individual student and the measurement of achievement for him or her, without reference to any other student. Later we comment on inferences that the consideration of scores for samples of individuals may make possible.

We wish to write a linear model for the observations which are taken to be the item scores for an individual. These item scores are assumed to be binary, i.e., either zero or one. The first model we write is:

$$Y_{ij} = \mu + a_j \tag{1}$$

where $\mu$ (mu) is an undefined constant that appears in the model largely for convenience when we later move to the consideration of data for a population of individuals, and $a_j$ characterizes individual $j$ within this population. It can be recognized immediately that this model implies that for every item $i$, individual $j$ secures the same item score $(\mu + a_j)$, and thus, when the data are binary, the vector of item scores for an individual consists only of all zeroes or all ones. Thus, model (1) makes an assumption about the item responses that nearly always can be shown to be false for some individual. Note, however, that it is possible to conceptualize a situation in which (1) might hold: Let the items define achievement in an esoteric subject matter that individuals never learn incidentally but only from formal instruction; then, uninstructed persons would be expected to be characterized by a vector of zero item responses. Also, let the formal instruction be such that an individual who is completely instructed knows all the items in the universe; then, instructed individuals would be expected to be characterized by a vector of item responses each equal to one. Since such situations are rare, we modify model (1) to permit the representation of other vectors of item responses.

---

*We ignore the problem of guessing that is associated with multiple choice items, and assume throughout that these are production items.

Model (2) may be written:

$$Y_{ij} = \mu + a_j + e_{ij}, \tag{2}$$

with $e_{ij}$ defined as a deviation score that results when $(\mu + a_j)$ is subtracted from the item score. Thus, for binary item scores for a given individual, $e_{ij}$ takes on only two values: $1 - (\mu + a_j)$ and $-(\mu + a_j)$. Note that when every $e_{ij}$ is zero, (2) becomes (1). Model (2) can now accommodate a vector with item response entries consisting of any pattern of zeroes and ones. It does this by taking the expected value of $e_{ij}$, over items, to be zero for individual $j$; as a consequence $(\mu + a_j)$ is the mean item score for individual $j$, and is a proportion $P_j$, say $0 < P_j < 1$. Note that $P_j$ is a parameter for individual $j$. It then follows that under model (2) an unbiased (and maximum likelihood) point estimate of $(\mu + a_j)$ is given for individual $j$ by the observed proportion correct of a random sample of $m$ items, and that the observed number correct for a fixed item sample size has a sampling distribution which is binomial with $P_j = (\mu + a_j)$ as the parameter and the number of items as the index. It also follows that for an item sample of fixed size an "error" can be defined as the mean of the $e_{ij}$ values for those items included in the sample for that individual $(M_{e \cdot j})$ and that this "error" also has a sampling distribution which is simply the linear transformation of the binomial distribution described above that results when the constant $(\mu + a_j)$ is subtracted from each observed proportion correct. Note that $(\mu + a_j)$ is a constant for any individual, but is not necessarily a constant across individuals.

An item sampling model of this sort is well known. (See Lord and Novick, 1968, pp.250-252). The unbiased point estimate of $(\mu + a_j)$ described above holds for a single individual regardless of variability in item characteristics such as difficulty so long as a random sample of items is used. This point may not be well understood. A "classic" confidence interval for $(\mu + a_j)$ may be developed on the basis of the observed proportion correct for m items, utilizing binomial distributions. Clopper and Pearson (1934) described such a procedure quite early; Walker and Lev (1953, Chapter 3) show the rationale and give illustrations of such confidence intervals. When the normal distribution is an adequate approximation to the binomial, procedures such as the two described by Hays (1973, section 9.26) may be used. Hays emphasizes the importance of a large sample size ($m$, here) for these procedures to be approximately correct. For large sample size, the binomial is well approximated by the normal distribution, especially for values of the parameter, $P_j$, in the range .20 to .80; when $P_j$ is very small or very large, the Poisson distribution gives a better approximation. Novick and Jackson (1974, sections 10-3 and 10-6) give methods employing the Poisson distribution which yield credible intervals for the posteriori distribution of the proportion correct. They suggest that these procedures are appropriate when the observed proportion correct is either .05 or less, or .95 or more, providing $m$ is greater than 10. The posteriori distribution is associated with Bayesian methods, and the credible interval is, very roughly, an analog of the classic confidence interval. Additional methods are available. Novick and Jackson (1974, section 10-4) discuss a normal approximation to the posteriori beta distribution that may be used to develop an interval that is slightly different from the two described by Hays. They also describe the arcsine transformation (using the Anscombe approximation) which yields, under rather general conditions, a variable which is approximately normally distributed with a known variance that is a function of $m$, the sample size. This transformation is often called a variance stabilizing transformation since the transformed variable has a variance which is no longer a function of the (unknown) parameter $P_j$. Novick and Jackson (1974, sections 10-1 and 10-6) describe the procedure, again in the context of a Bayesian analysis, which yields a credible interval for $P_j$ derived from the posteriori distribution. Note that in section 10-1, Novick and Jackson illustrate the use of a uniform prior distribution and in section 10-6 the use of an indifference prior. Finally, "direct" Bayesian methods use a beta distribution as a prior, and this, together with the binomial distribution, gives a posteriori beta distribution; the parameters of the two beta distributions are not the same, of course. One can then construct a credible interval either from tables of the appropriate beta distribution, which are given by Novick and Jackson (1974, Table A 14) or from tables of the F distribution. (See Hays, 1973, section 19.12, for an illustration, and Novick and Jackson, 1974, pp. 124-125 for an explanation.) This direct method for binary data is discussed in detail in Chapter 5 of Novick and Jackson (1974).

The model given in (2) accommodates any vector of zeroes and ones as the responses to a universe of items by individual $j$. A random sample of these responses then has a mean ($x/m$, say, where $x$ is the number correct) which is a point estimate of the unknown $P_j = (\mu + a_j)$, or proportion correct of the item universe. It also is true that $x/m$ is the mode of the Bayesian posteriori distribution when a beta distribution with parameters 1 and 1 (the uniform beta) is taken as the prior; this is related to the fact that $x/m$ is a maximum likelihood estimator of $P_j$. We present in Table 1 values for an interval constructed in several different ways as an illustration of these methods; the marked similarity of the results is in part a function of taking the observed proportion to be one-half.

Model (2) is a sampling error model for which there is no measurement error in the sense of specific unreliability, i.e., variation over parallel replications. Instead, $e_{ij}$ is what Lord and Novick (1968, Chapter 8) call a generic error. The variance of the $e_{ij}$ for a given individual

**Table 1**
**95% Confidence and Credible Intervals for $P_j$ Given that 10 of 20 Items are Answered Correctly**

| Method | Interval | |
|---|---|---|
| | Low | High |
| Clopper and Pearson (1934) Graphs | .27 | .73 |
| Walker and Lev (1953) Normal Approximation, Formula 3.3 | .30 | .70 |
| Hays (1973) Normal Approximation Formula 9.26.1 | .30 | .70 |
| Hays (1973) Normal Approximation Formula 9.26.2 | .28 | .73 |
| Novick and Jackson (1974) Normal Approximation, section 10-6 | .29 | .71 |
| Novick and Jackson (1974) Arcsine Normal Approximation, section 10-6 | .29 | .71 |
| Novick and Jackson (1974) Direct beta with indifference prior, section 10-6 | .29 | .71 |
| Novick and Jackson (1974) Direct beta with uniform prior, section 10-1 | .30 | .70 |

is the individual generic error variance; it is simply the expected value over items of $e_{ij}^2$ for a fixed $j$ since the expected value of $e_{ij}$ over items is zero. This expected value of $e_{ij}^2$ can be shown to equal $P_jQ_j$, where $Q_j = (1 - P_j)$ and thus, is a function of the individual parameter $P_j$. It follows that if two individuals are at different levels of instruction, with $P_j \neq P_k$, their individual generic variances will differ. (An exception occurs when $P_j = 1 - P_k$, since then $P_jQ_j = P_kQ_k$, as for example with $P_j = .4$ and $P_k = .6$) It is possible to make an unbiased estimate of this generic error variance for any individual from the responses to two or more randomly selected items. For two items, the estimate is simply one-half the squared difference between scores on the two items. For binary items, this squared difference takes values only of zero or one. For m items, the unbiased estimate is simply $mp_jq_j/(m - 1)$ where $p_j$ is the observed proportion correct for the m items.

The individual generic error variance can be averaged over a population of individuals to yield the group generic error variance. Assume model (2) for each individual in a population. Then the group generic error variance is the expected value of $e_{ij}^2$ over both items and individuals. With more than one individual we have unmatched sample data when each individual responds to a different random sample of items, and matched sample data when each individual responds to the same random sample of items.

For either case, an unbiased estimate of the group generic error variance can be secured by averaging the unbiased estimates of the individual generic error variance under the restriction that $m$, the number of items, is constant for each individual. For unmatched data, this estimate is the analog of the observed mean square within individuals in a one-way analysis of variance framework where individuals play the role conventionally assigned to treatments in this anova framework and items are nested within individuals. It is also important conceptually to recognize that homogeneity of individual generic error variance generally does not hold, and that this observed mean square has as an expected value an average variance and not simply an assumed constant variance. We will consider matched data later.

The one-way analysis of variance framework has a second observed mean-square, that between individuals. We assume that the J individuals entering the sample of individuals are randomly drawn and that each individual responds to a different random sample of $m$ items (unmatched data). The structure is like that of a one-way random anova, and the expected value of the mean square between persons has three components, one of which is null. These are:

$$mE \sum_j \frac{(a_j - M_{\bullet})^2}{J - 1}$$

$$mE \sum_j \frac{(M_{\bullet \cdot j} - M_{\bullet \cdot \cdot})^2}{J - 1}$$

$$2mE \sum_j \frac{(a_j - M_{\bullet})(M_{\bullet \cdot j} - M_{\bullet \cdot \cdot})}{J - 1}$$

This first expected value is simply $m$ times the population variance of the $a_j$ values. The second expected value is the group generic error variance for the population. The third term should be understood as an expectation over items within individuals and an expectation over individuals. Under these conditions, this term is zero even though the population covariance of the $a_j$ and the $e_{ij}$ for any one item is not necessarily zero; Lord and Novick (1968 p. 182) show this using a double expectation. This anova structure therefore gives, as the expected value for the mean square between persons, $m$ times the population variance of the $a_j$ plus the population group generic error variance; the expected value for the mean square within persons is the population group generic error variance. The observed values of the anova therefore can provide an unbiased estimate of the population variance of the $a_j$. We are concerned here with estimation, and not with tests of significance. Further, we certainly do not imply that a variance-ratio test on the observed mean squares is justified by these remarks about unbiased estimates; additional distribution assumptions clearly are needed for such tests. We also believe that $m$ should be a constant for each individual in this anova framework, and we interpret Hayes (1973, pp. 529-530) as supporting this position.

Next, let us postulate a row-by-column item response matrix (indefinitely large in both directions) with rows designating individuals ($j$) and columns designating items ($i$). Such a matrix can accomodate indefinitely many item responses of indefinitely many individuals. Centering this matrix only by rows gives the $e_{ij}$ of model (2); now let us conceptually double-center this matrix in order to define two parts of $e_{ij}$. One way to conceptualize the double centering is to regard it as subtracting both the row and column mean from any cell entry and then adding back in the grand mean for the whole matrix. We then specify as a model:

$$Y_{ij} = \mu + a_j + b_i + f_{ij}. \tag{3}$$

$Y_{ij}$ is the binary score (zero or one) of individual $j$ on item $i$; $\mu$ and $a_j$ have the same definitions as for model (2). The double centering splits $e_{ij}$ into the two parts $(b_i + f_{ij})$, where $b_i$ is a constant for individuals but has an expected value of zero over items. We may describe $b_i$ as the relative bias of item $i$, and $(\mu + b_i)$ as the (normative) difficulty of item $i$. When every $b_i$ is zero and thus, items are of equal difficulty, (3) becomes (2) with $f_{ij} = e_{ij}$. We point out here that $f_{ij}$ has an expected value of zero over both items and individuals, and that the generic error of measurement is made up of the relative bias of the item $(b_i)$ plus this residual $(f_{ij})$.

We can identify (3) with the Cornfield and Tukey (1956, section 8, esp.) pigeonhole model in which there is a single value in each cell or pigeonhole. They describe this as an additive model for a bisample of rows and columns drawn from an arbitrary row-by-column matrix of constants (p. 918). The constants are the item scores for the individuals. They also point out that the $f_{ij}$ terms are not interactions in the sense of being well-defined functions of the row and column parameters, since there can be row and column combinations for which these parameters are equal, but the $f_{ij}$ are not equal. Instead, the $f_{ij}$ elements of (3) are defined strictly as the "residuals" given by double-centering the complete matrix, and they are not the "errors" that ordinarily would be defined by considering fluctuation over occasions or replications of the item scores.

Let us define $(\mu + a_j)$ as the domain score (proportion correct) for an individual and $(b_i + f_{ij})$ as the error of measurement as before. Then the variance of a fixed item is the expected value over individuals of $(a_j + f_{ij})^2$ which is the expected value of the sum $(a_j^2 + f_{ij}^2 + 2a_jf_{ij})$. Since $a_jf_{ij}$ does not necessarily have an expected value of zero, it follows that the variance of an item has three components. One of these is the expected value of $a_j^2$ and is the domain score variance, since $\mu$ is a constant over the whole matrix. A second is the expected value of $f_{ij}^2$ which is the variable portion (over individuals) of the error of measurement. We can conclude then that the observed score variance (for any item) is not equal to the sum of domain—or "true"—score variance and error variance. We also can deduce that the residuals for two different items may not necessarily be uncorrelated in the population of individuals. Consider a finite universe of items. The population variance-covariance matrix for the residuals (with, of course, variance and covariance defined over individuals) necessarily sums to zero both ways since the expected value of the $f_{ij}$ is zero over items. Thus, when the diagonal terms of this variance-covariance matrix (the variances) are positive, as they always should be, the off-diagonal terms (the covariances) in any array (row or column) cannot all be zero, and thus, the intercorrelations of these residuals cannot all be zero. Since the $f_{ij}$ make up the variable (over individuals) portion of the generic error of measurement, it follows that the errors of measurement for different items are not necessarily uncorrelated. Again, considering a finite universe of items, we observe that with model (3) it is impossible for the following two conditions to hold simultaneously:

(a) an observed proportion right for $m$ items is an unbiased estimate of the domain proportion right for every individual;

(b) errors of measurement are uncorrelated in the population of individuals for all pairs of items.

It is important to stress this point, since for achievement tests it often is reasonable to regard the item domain as finite, even though it may be quite large. For example, items consisting of five two-digit numbers to be added make up a domain of $10^{10}$ items ($9^{10}$ if zero is excluded; a smaller number if position in the column is considered); this is a fairly "large" finite domain. For such a domain, an item sampling model would be attractive because of the possibility of unbiased estimation of the domain proportion right for any individual even though one must accept correlated errors.

We regard (3) as a appropriate model for what we call the single-pass matrix. The notion is that we can imagine or conceptualize each individual's attempting each of the items and that the results of this single pass over all items by all individuals are contained in the matrix of zeroes and ones. We would argue that in many instructional situations it is both realistic and appropriate to regard the data of this single-pass matrix as the data of interest. The question that those managing the instruction often wish to answer is: How well can the individuals perform on this universe of items at this time? By using an item sampling principle relevant estimates for particular individuals can be made. By using a matrix sampling principle, which yields matched data, estimates can be made of the group generic error variance, of item bias, and of the covariance of the observed score on an item with the generic "true" score, which is the domain proportion right.

The term $b_i$ in (3) has already been identified as the relative bias of an item. This bias may be estimated by fixing that item and then comparing the observed mean for that item with the observed mean proportion correct of a randomly selected set of items, where the set does not include the fixed item. For only two items, one of which is fixed and the other of which is chosen randomly from an infinite universe of items, the estimate is simply the difficulty of the fixed item minus the difficulty of the randomly sampled item. In the next chapter, we consider this estimate as an index of departure from equal difficulties. It also is possible, for a fixed item, to estimate the covariance of the observed score on this item with the generic "true" score; the estimate is simply the average of the covariances between this item and a random sample of distinct items (using $n - 1$ as a divisor for each covariance). Lord and Novick (1968, pp. 190–191) regard such a covariance as providing important information about an item since, in this model, this covariance is not generally equal to the "true" score variance as it would be for a model of strictly parallel tests. The data for a matrix sample may be treated in an analysis of variance framework, yielding observed mean squares for items, for individuals, and for residuals. We have already pointed out that this third mean square is an unbiased estimate of the group generic error variance. Lord and Novick (1968, section 11.5) show how the variance (over individuals) of the domain proportion right score may be estimated from a matrix sample. Sirotnik (1970) gives a similar demonstration. In a matrix sample (matched study), when the same two binary items are used, the estimate of the group generic error variance based on responses to the two items is simply $(b + c)/2n$ where $(b + c)/n$ is the proportion of individuals who make inconsistent responses to the two items (0, 1 or 1, 0). This type of estimate plays a role in the work reported in Chapters 4 and 5.

Important characteristics of model (3) may be summarized in this fashion. The model provides a domain proportion correct $(\mu + a_j)$, a measure of relative bias for the item $(b_i)$, and a "residual" $(f_{ij})$. This residual is the variable (over individuals) portion of the individual generic error variance, and is not in general uncorrelated with any other residual or with $a_j$ which is the variable (over individuals) portion of the generic "true" score. Since generic "true" and "error" components of any item score are not uncorrelated in a population of individuals, there is no unique generic reliability coefficient for an item or for a "test" made up of a sum of items. Group generic error variance, generic "true" score variance, and the covariance of observed score with generic "true" score can all be estimated, but observed score variance is not in general equal to the sum of group generic error variance and generic "true" score variance for an item. Thus, for one item or for a test of several items, the familiar ratio of "true" variance to observed variance is not necessarily equal to one minus the ratio of "error" variance to observed variance, and neither of these is necessarily equal to the squared correlation of "true" score with observed score.

These characteristics indicate that there are limitations associated with the item sampling model in that conventional interpretations of such item achievement data may not be appropriate or adequate. If, as we maintain, the single-pass matrix is a realistic model for certain types of achievement testing, the problem of appropriate interpretation and analysis deserves further examination. An attempt to do this is represented by work that appears in the following chapters.

We now consider briefly a second concern about the variability of item scores. In our discussion of models (2) and (3), we examined variability across an item domain for a population of individuals; to use a conventional term, we were concerned with item equivalence. We now turn to a concern about variability over replications which is reasonably well described as a concern with item stability.

It appears to be conventional to regard replications as strictly parallel measures, at least in theory. If replications are strictly parallel, then the error of measurement may be labeled specific, rather than generic, and several important conditions hold generally for model (4), which is written for a fixed item:

$$Y_{jr} = \mu + a_j + e_{jr} \tag{4}$$

($Y_{jr}$ is the binary score on the fixed item for person $j$ on replication $r$, $\mu$ is the mean score for this item over both replications and individuals, $a_j$ is the mean score of person $j$ over replications, and $e_{jr}$ is the specific error.) Under parallelism, conditions that hold are: (1) The expected value of the individual specific errors over replications is zero, and so we have unbiased estimation of the individual's item score. (2) For any replication, the observed variance is equal to the sum of "true" and error variance, and the error component $e_{jr}$ is uncorrelated with the "true" component ($\mu + a_j$). (3) Further, the error components for different replications are uncorrelated with each other, and for any two replications the covariance of "true" plus error components equals the "true" variance component. This fits the well-known "classic" model (Lord and Novick, 1968, Chapter 3) Thus, when strict parallelism holds for replications, estimates of specific reliability are straight-forward and give a sufficient characterization of the item or sets of items as a measuring instrument.

We now wish to develop for model (4) an argument similar to one developed for model (3). First, let us assume that replications have the same population mean (over individuals) and thus, assume that it is not necessary to separate $e_{jr}$ into two parts. (Recall that we separated the $e_{ij}$ of model (2) into two parts, giving us model (3), on the grounds that the items are likely to differ in difficulty.) Next, let us consider a finite number of replications for a population of individuals and construct the variance-covariance matrix of the $e_{jr}$ terms. Unbiased estimation of an individual's item score in model (4) requires that the expected value of $e_{jr}$ over replications equals zero. If so, it follows that this variance-covariance matrix must sum to zero both ways and that the covariances cannot all be zero so long as the $e_{jr}$ are real numbers and thus, have non-negative variances. We therefore conclude that for a finite number of replications we cannot have, simultaneously, unbiased estimation of an individual's item score and linear experimental independence (zero correlation) of the specific error terms $e_{jr}$. In other words, for a finite number of replications, assumptions 3.1.2 and 3.1.4 of the classic model (Lord and Novick, 1968, p. 56) cannot hold simultaneously for all errors.

We now conjecture that a similar conclusion holds for a denumerably infinite set of replications. It seems clear that replications must be countable and thus, can only be denumerably infinite. Expected values exist for a denumerably infinite set (consider for example the mean of an infinite number of observations on a population of individuals), and so we conjecture that it is meaningful to conclude that zero is the expected value for any denumerably infinite array of the variance-covariance matrix of the $e_{jr}$. If this holds, then the expected value of only the covariances of any replication error, $e_{jr}$, with the remaining errors cannot be zero, and thus, the correlations cannot all be zero. We are careful to label this a conjecture, and we would welcome commentary concerning it. If our conjecture is correct, the notion of specific reliability and its estimation may need reexamination.

In our discussion of data models, we have drawn a number of points from the work of Lord and Novick (1968), particularly from their Chapter 8. They point out that earlier work by Cronbach and his associates develops the notion and advocates the use of a generic true score; much of this work is now available in a single volume (Cronbach, *et al*, 1972). Recently, Cardinet *et al* (1976) suggested applications of generalizability theory to educational measurements, making extensive use of intraclass correlation coefficients.

We merely state again that such coefficients are somewhat ambiguous when the "true" and error components of a measurement are not linearly experimentally independent; Lord and Novick (1968, Section 11.5) regard such a variance ratio as inadequate for describing the relation of observed score to (generic) "true" score under these conditions.

We close this chapter by emphasizing limitations which characterize it. We have not mentioned what are called fixed effects in the analysis of variance; the Cronbach and Cardinet references may be consulted for information about possible roles of such effects in measurement situations. We see the random selection of items from a domain as a critical requirement for the type of achievement testing we are concerned with.

For one individual, this permits a meaningful inference about his domain score. By also considering the random sampling of individuals, we are able to employ the one-way random anova framework for the estimation of individual and group generic error and generic "true" score variance; this is an unmatched data study. We have emphasized the analysis of matched data in our discussions of models (3) and (4), and we continue this emphasis in the next chapter. We have given no attention to the problem of sampling replications. It may be that the fact that replications occur in time makes their sampling a special problem that cannot be solved by a model like an urn model.

Finally, we have not attempted to bring models (3) and (4) together into a single design. We see that this might be done in two different ways: One would be to regard replications as nested within the pigeonholes of model (3); for any fixed individual and item, the propensity distribution would describe these replications. Another would be to set up a three-way, completely crossed design, with individuals, items, and replications as the main effects or facets; Cardinet *et al* (1976) illustrate a similar three-way design with moments, objects, and subjects as the facets. We recognize that it is conventional to label certain variance components "interactions," but we cite Cornfield and Tukey (1956) again to call attention to the distinction between interactions (which are functions of the row, column, and or slice parameters) and residuals in the pigeonhole model. Such a pigeonhole model can, of course, be three dimensional. The relative merits of the nested and crossed designs probably need further analysis.

## Chapter 3
# CONVENTIONAL SIGNIFICANCE TESTS AND INDICES OF AGREEMENT OR ASSOCIATION

CHESTER W. HARRIS
ANDREA PASTOROK PEARLMAN

Throughout the remainder of these papers, we shall be concerned with data that can be summarized meaningfully in a two-by-two or fourfold table. We can specify several situations or types of study for which such a data summary is an appropriate one. Given a sample of $n$ individuals, we may have data which provides two distinct dichotomizations of the $n$ individuals. A familiar example occurs when we have two binary item scores (0, 1) for each individual; it follows that focusing attention on one of the items provides a dichotomization of the individuals into two groups, and focusing on the other item provides a second dichotomization of the same individuals. Considering both items jointly yields four groups: those who answer both items correctly, those who answer both items incorrectly, those who answer Item A correctly and Item B incorrectly, and those who answer Item A incorrectly and Item B correctly. We can count the number of persons in each of these four groups and thus create the following table:

**Display A (Observations)**

| | | Item B | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Item A | 1 | $a$ | $b$ | $a + b$ |
| | 0 | $c$ | $d$ | $c + d$ |
| | | $a + c$ | $b + d$ | $n$ |

Here $a$, $b$, $c$, and $d$ are the observed frequencies of the four categories of individuals, and these observed frequencies sum to $n$. The headings 1 and 0 designate the item scores. The marginals $(a + b)$, $(c + d)$, $(a + c)$ and $(b + d)$ are the frequencies observed for the headings 1 and 0 for the two observations. Such a display also can accommodate the data given by a study in which the same item is administered on two separate occasions to the same sample of individuals. For this type of study, rows and columns designate not different items, but two different occasions, and may be labeled First Administration and Second Administration as in the study of the sensitivity of an item to instruction.

**Display B (Observations)**

| | | Second Administration | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| First Administration | 1 | $a$ | $b$ | $a + b$ |
| | 0 | $c$ | $d$ | $c + d$ |
| | | $a + c$ | $b + d$ | $n$ |

Frequency data may be transformed to proportions. By using $n$ as a divisor for each frequency in the two-by-two table, we create a new display of sample proportions.

**Display A (Sample Proportions)**

| | | Item B | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Item A | 1 | $p_{11}$ | $p_{10}$ | $p_{1.}$ |
| | 0 | $p_{01}$ | $p_{00}$ | $p_{0.}$ |
| | | $p_{.1}$ | $p_{.0}$ | 1 |

Note that $(a + b)/n$ gives $p_{1.}$, or the observed proportion correct for Item A, and that $p_{10}$ gives the proportion of persons for whom Item A is coded 1 and Item B is coded 0. Marginal proportions are indicated by using a dot to designate the observation over which the sum is computed. In some of the discussion, we shall be concerned about the temporal sequence of observations and will regard a statistic like $p_{10}$ as indicating the sample proportion for which the First Administration is coded 1 and the Second Administration is coded 0.

**Display B (Sample Proportions)**

| | | Second Administration | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| First Administration | 1 | $p_{11}$ | $p_{10}$ | $p_{1.}$ |
| | 0 | $p_{01}$ | $p_{00}$ | $p_{0.}$ |
| | | $p_{.1}$ | $p_{.0}$ | 1 |

We can now write the population proportions or parameters that correspond to these sample values.

**Display A (Parameters)**

| | | Item B | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Item A | 1 | $P_{11}$ | $P_{10}$ | $P_{1.}$ |
| | 0 | $P_{01}$ | $P_{00}$ | $P_{0.}$ |
| | | $P_{.1}$ | $P_{.0}$ | 1 |

Here $P_{1.}$ is the proportion of individuals in the *population* who answer Item A correctly. A similar display can be constructed for first and second administrations.

**Display B (Parameters)**

| | | Second Administration | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| First Administration | 1 | $P_{11}$ | $P_{10}$ | $P_{1.}$ |
| | 0 | $P_{01}$ | $P_{00}$ | $P_{0.}$ |
| | | $P_{.1}$ | $P_{.0}$ | 1 |

In specifying various hypotheses that may be tested, we refer explicitly to parameters. In describing estimation procedures and the sample values employed in testing hypotheses, we refer to the observations or to the sample proportions.

The headings 1 and 0 for either the rows or the columns as discussed thus far represent categorizations of the sample or of the population based on performance on a single test item. However, we also want to consider significance tests and estimation procedures when one of the two categorizations is made on a slightly different basis. For example, the headings 1 and 0 for either the rows or the columns may also represent categorizations of the sample or the population based on instructional history, such as Instructed and Not Instructed. These three displays (Display C) would be relevant in a prospective study of sensitivity of an item to instruction. We discuss this type of study later and show how it differs from a study in which the same item is administered prior to and following instruction, a design for which the First Administration, Second Administration Display (Display B) is relevant.

**Display C (Observations)**

| | Item | | |
|---|---|---|---|
| | 1 | 0 | |
| Instructed 1 | $a$ | $b$ | $a + b$ |
| Not Instructed 0 | $c$ | $d$ | $c + d$ |
| | $a + c$ | $b + d$ | $n$ |

**Display C (Sample Proportions)**

| | Item | | |
|---|---|---|---|
| | 1 | 0 | |
| Instructed 1 | $p_{11}$ | $p_{10}$ | $p_{1.}$ |
| Not Instructed 0 | $p_{01}$ | $p_{00}$ | $p_{0.}$ |
| | $p_{.1}$ | $p_{.0}$ | 1 |

**Display C (Paramaters)**

|  | Item |  |  |
|---|---|---|---|
|  | 1 | 0 |  |
| Instructed 1 | $P_{11}$ | $P_{10}$ | $P_{1.}$ |
| Not Instructed 0 | $P_{01}$ | $P_{00}$ | $P_{0.}$ |
|  | $P_{.1}$ | $P_{.0}$ | 1 |

Another method of arranging the data for two items is given by Display D (Observations), which is a modification of Display A.

**Display D (Observations)**

|  |  | Item B |  |  |
|---|---|---|---|---|
|  |  | Same | Different |  |
| Item A | 1 | $a$ | $b$ | $a + b$ |
|  | 0 | $d$ | $c$ | $c + d$ |
|  |  | $a + d$ | $b + c$ | $n$ |

This arrangement focuses attention on the consistency or inconsistency of the responses to the two items, with Same meaning both responses correct (all $a$) or both responses incorrect (all $d$). The sample proportions and parameters are:

**Display D (Sample Proportions)**

|  |  | Item B |  |  |
|---|---|---|---|---|
|  |  | Same | Different |  |
| Item A | 1 | $p_{11}$ | $p_{10}$ | $p_{1.}$ |
|  | 0 | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
|  |  | $p_{11} + p_{00}$ | $p_{10} + p_{01}$ | 1 |

**Display D (Parameters)**

|  |  | Item B |  |  |
|---|---|---|---|---|
|  |  | Same | Different |  |
| Item A | 1 | $P_{11}$ | $P_{10}$ | $P_{1.}$ |
|  | 0 | $P_{00}$ | $P_{01}$ | $P_{0.}$ |
|  |  | $P_{11} + P_{00}$ | $P_{10} + P_{01}$ | 1 |

Whenever we have data for the same item administered on two occasions it is possible to modify Display B to emphasize the correspondence of the response on one occasion to the response on the other occasion. For example, we may create this display:

**Display E (Observations)**

| | | Second Administration | | |
|---|---|---|---|---|
| | | Same | Different | |
| First Administration | 1 | $a$ | $b$ | $a + b$ |
| | 0 | $d$ | $c$ | $c + d$ |
| | | $a + d$ | $b + c$ | $n$ |

The sample proportions and the parameters are:

**Display E (Sample Proportions)**

| | | Second Administration | | |
|---|---|---|---|---|
| | | Same | Different | |
| First Administration | 1 | $p_{11}$ | $p_{10}$ | $p_{1.}$ |
| | 0 | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
| | | $p_{11} + p_{00}$ | $p_{10} + p_{01}$ | 1 |

**Display E (Parameters)**

| | | Second Administration | | |
|---|---|---|---|---|
| | | Same | Different | |
| First Administration | 1 | $P_{11}$ | $P_{10}$ | $P_{1.}$ |
| | 0 | $P_{00}$ | $P_{01}$ | $P_{0.}$ |
| | | $P_{11} + P_{00}$ | $P_{10} + P_{01}$ | 1 |

Note that duals of these three displays exist if we take the second administration as the referent and code the result of the first administration as Same or Different; these duals are transposes of the matrices.

*The Chi Square Test of Significance.* It is possible to distinguish different significance tests for these two-by-two tables. The best known of these probably is the conventional chi square test, which is an approximate test of independence, with independence referring to a situation in which the population values within the four cells are perfectly predictable from the population marginal values. For a two-by-two table, the conventional (uncorrected) chi square sample value is computed as

$$\frac{n(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)}.$$

Here the chi square test of independence (which is also a test of the hypothesis that $P_{11}/P_{.1} = P_{10}/P_{.0}$ or, alternatively, that $P_{11}/P_{1.} = P_{01}/P_{0.}$) has one degree of freedom since $v = RC - 1 - (C - 1) - (R - 1) = (R - 1)(C - 1) = 1$, where $R$ designates the number of rows and $C$ the number of columns. For a fourfold table, which has degrees of freedom equal to unity, the chi square measure is also the Cramér measure which

Goodman and Kruskal (1954) regard as giving a better norming than the contingency coefficient. They also warn that:

> "The fact that an excellent test of independence may be based on $\chi^2$ does not at all mean that $\chi^2$, or some simple function of it is an appropriate *measure* of degree of association." (Goodman & Kruskal, 1954, p. 740)

When degrees of freedom equal unity, Yates's correction for continuity may be employed with the chi square test. The size of chi square is reduced by changing the frequency in each cell by .5 yet maintaining the marginal totals. Thus, the chi square sample value adjusted by Yates's correction is computed as

$$\frac{n(|ad - bc| - n/2)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

The Fisher exact test for a two-by-two table is, of course, always available rather than the chi square test which is an approximation to the randomization test represented by the Fisher method. For large samples, the Fisher exact test is laborious to compute. Hays (1973, p. 738) observes that Fisher's exact test computes the exact probability "for a sample's showing as much or more evidence for association than that obtained, given only the operation of chance." It is essentially a one-tailed test which tests the null hypothesis that any apparent association in the table is determined by chance. When the actual table obtained falls among those unlikely to occur over all random assignments of the subjects in columns to the rows, or of subjects in rows to the columns, the null hypothesis is rejected in favor of one which says association exists beyond chance expectation.

Garside and Mack (1976) investigated the Type I error rate in the homogeneity case (two of the marginal totals are fixed in advance as might be true of Display C) of the two-by-two contingency table. For the homogeneity case, Garside (1971) and Boschloo (1970) published tables of corrections which purported to give better results than either the Fisher exact test or $\chi^2$ with Yates's correction since their Type I error rates were in general closer to $\alpha$ and did not exceed $\alpha$. Calculations of the actual Type I error probabilities were computed for five tests: the uncorrected $\chi^2$ test, $\chi^2$ with Yates's correction, $\chi^2$ with Garside's correction, Fisher's exact test, and the Boschloo modification of Fisher's test. The results of the Garside and Mack (1976, p. 18) investigation,

> show that, in general, Boschloo's and Garside's error probabilities are very similar and much closer to $\alpha$ than Fisher's and Yates' errors (which are very similar to each other). . . . The uncorrected chi square test gives actual error probabilities which usually exceed $\alpha$ for some values. . . .

We now wish to consider the use of the chi square test in studies that can be represented by Displays A, B and C. For Display A (Parameters), independence means that $P_{11} = P_{1.}P_{.1}$, $P_{10} = P_{1.}P_{.0}$, $P_{01} = P_{0.}P_{.1}$, and $P_{00} = P_{0.}P_{.0}$. When independence holds, it will also be true that, e.g., $P_{11}/P_{1.} = P_{.1} = P_{01}/P_{0.}$; in other words, the conditional probabilities $P_{11}/P_{1.}$ and $P_{01}/P_{0.}$ are equal. This would mean that the same proportion of the population answering Item A correctly and the population answering Item A incorrectly would answer Item B correctly. (Independence also means that $P_{10}/P_{1.} = P_{00}/P_{0.}$.) If the hypothesis of independence is rejected as a result of the $\chi^2$ test on the sample values, we may conclude that the responses to the two items are associated (not independent); we do not, however, have in the probability associated with the sample $\chi^2$ value a measure of the degree of association. This is a first point about the use of $\chi^2$ with Displays A, B and C.

In the study of respones to two different items (Display A), it may be informative to test the hypothesis of independence; however, when the two items are drawn from the same item domain then we may quite reasonably expect to reject this null hypothesis and thus, we may learn very little from a significant test result. In a situation such as this, it is a measure of association (or possibly of agreement) that could be more informative. Later in this chapter, we shall review some conventional measures of this type and comment on their utility. Also, in Chapter 5 we consider some less familiar approaches to estimating agreement or association.

Next, let us consider the use of the $\chi^2$ test with the administration of the same item on two distinct occasions (Display B) that are closely related in time. We assume that during this short period of time there is no major change in the performance of the individuals, and that the changes in response that do occur are evidence of lack of stability, or specific reliability in Lord and Novick's terms (1968, Chapter 7). Under these conditions, it is hardly reasonable to expect independence of the responses on the two occasions, and consequently the test of independence would seldom be informative. Instead, the outcome of the $\chi^2$ test in this situation is, almost certainly, merely a reflection of the power of the test. The test of independence for Display B may be stated as:

$$H_0: \quad \frac{P_{11}}{P_{1.}} = \frac{P_{01}}{P_{0.}}$$

$$H_1: \quad \frac{P_{11}}{P_{1.}} \neq \frac{P_{01}}{P_{0.}}$$

(This is the same as for Display A.) Rejecting $H_0$ in favor of $H_1$ leads to the conclusion that the conditional probability of answering the item correctly on the second occasion is different for two populations: those who answer it correctly on the first administration and those who answer it incorrectly on the first administration. This is what is expected;

however, the expectation goes further, since in this situation we would be somewhat astonished to find $P_{01}$ greater than $P_{11}$, even though this is consistent with rejecting $H_0$ in favor of $H_1$. In other words, we expect a positive relation between responses to the item on the two occasions, rather than a negative relation; but either relation is consistent with a rejection of the hypothesis of independence. We conclude that the $\chi^2$ test often is not informative in the case of Display B as well as the case of Display A, and that, for both cases, measures of association (including the direction of the association) or agreement are more informative. We return to this topic later in this chapter.

Now let us consider Display C and the case of the response to an item by two populations: students who have been instructed and students who have not been instructed. When the item is relevant to the instruction, we clearly expect $P_{11}/P_{1.}$ to be greater than $P_{01}/P_{0.}$, and so we expect to reject the hypothesis of independence. In this situation, like the earlier one, the outcome of the $\chi^2$ test primarily reflects the power of the test. Also, rejecting the hypothesis of independence in itself does not answer the critical questions of the direction of the association and of the strength of the association. These comments about the $\chi^2$ test merely set the stage for further discussion of the study of item sensitivity to instruction later.

*The McNemar Test of Significance.* A second (approximate) significance test is given by the McNemar test, which tests the hypothesis that the population proportions corresponding to $(a + b)/n$ and $(a + c)/n$ or $(c + d)/n$ and $(b + d)/n$ are equal. This is a test of correlated proportions, where each sample proportion involves some of the same observations; thus, the two samples are not independent. The uncorrected McNemar statistic is calculated as $(b - c)^2/(b + c)$ and referred to the chi square table with one degree of freedom.

There also is an exact test which employs the binomial distribution; Hays (1973, p. 741) shows how this can be developed. In terms of parameters, the hypothesis may be stated as

$$H_0: \quad P_{.1} = P_{1.} \qquad\qquad H_0: \quad P_{.0} = P_{0.}$$

or

$$H_1: \quad P_{.1} \neq P_{1.} \qquad\qquad H_1: \quad P_{.0} \neq P_{0.}$$

When $H_0$ is true, the probability of a given sample with cell frequencies $b$ and $c$ is $\binom{b+c}{b}(.5)^{b+c}$. To carry out the exact test, Hays lets $g$ equal the smaller of the two frequencies, $b$ or $c$, and then takes the sum of probabilities $2 \sum_{k=0}^{g} \binom{b+c}{k}(.5)^{b+c}$. For a two-tailed test, the null hypothesis may be rejected if this number is less than or equal to the value chosen for $\alpha$. Hays states that, when $n$ is relatively large, the exact probability may be approximated by use of a corrected chi square with one degree of freedom where

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}.$$

In Chapter 4, we examine some power characteristics of the uncorrected test statistic.

For Display A or Display B, the McNemar test is a test of the hypothesis of identical population item difficulties. It may be regarded as a test of exact symmetry; later, we consider a test of relative symmetry. Consider Display A (Parameters). The hypothesis being tested by the McNemar test does not duplicate the test of independence; instead, it is a test of the exact symmetry of the marginals for the two items. This can be an informative test, for example, in a situation in which one expects two different items (but possibly items from the same item domain) to have been learned equally well during instruction; a significant outcome would indicate that this expectation had not been borne out. If, in contrast, the expectation is that the item difficulties are different, then this test is inappropriate.

Display B (Parameters) refers to a different type of situation—that of the same item being administered twice. We can identify two cases that fit this situation; in one of these the McNemar test would be meaningful and in the other it would not. In a study of stability, as described above, one of the expectations is that the difficulty of the item remains the same over the two administrations that are closely related in time. That is, in a sense, a first level expectation, since if this expectation is false we clearly do not have stability, but if it is true we still have an additional expectation about agreement or association to consider and possibly test. We can use the McNemar test to check this first level expectation, but we must consider the power of the test in interpreting the results. See Chapter 4 for a suggested procedure of this kind. A second case is that of administering the item before and after instruction; here the expectation is that instruction will alter the difficulty ("easiness") of the item, and consequently $H_0$: $P_{1.} = P_{.1}$ is expected to be false. In this situation, use of the McNemar test is inappropriate or, at best uninformative.

We conclude this section by pointing out that the McNemar test seems useless for Displays C, D, or E.

*A Test of Relative Symmetry.* We now wish to examine Displays D and E which were constructed from Display A and Display B by relabeling the column categories as Same and Different and rearranging the frequencies (or proportions) in the second row. A dual exists when we focus on rows rather than columns. A test of relative symmetry would test the hypothesis that the population proportion of consistent responses is the same for persons in the two row categories. The hypothesis may be stated:

$$H_0: \quad \frac{P_{11}}{P_{1.}} = \frac{P_{00}}{P_{0.}}$$

$$H_1: \quad \frac{P_{11}}{P_{1.}} \neq \frac{P_{00}}{P_{0.}}$$

For Display D, when $H_0$ is true, the conditional probability of answering the two items the same way is the same for those who answered Item A correctly and for those who answered it incorrectly. For Display E, $H_0$ true means that the conditional probability of answering the item the same way twice is the same for those who were correct and those who were incorrect on the first administration of the item. For Display E, when systematic instruction has intervened between the two administrations (a pre- and post-test study), the conditional probability of responding correctly on the second administration given an incorrect response on the first administration is $P_{01}/P_{0.}$. Analogously, the conditional probability of responding incorrectly on the second administration given a correct response on the first administration is $P_{10}/P_{1.}$. The difference between these two conditional probabilities might be taken as an index or measure of the amount of learning. Relative symmetry implies that this index $[P_{01}/P_{0.} - P_{10}/P_{1.}]$ is zero. (This is true because the hypothesis of relative symmetry may be expressed either as $P_{11}/P_{1.} = P_{00}/P_{0.}$ or $P_{10}/P_{1.} = P_{01}/P_{0.}$.) Bishop *et al* (1975, section 7.1.2) discuss relative symmetry under the heading "Models for measuring change."

A test of relative symmetry would be made by computing the (uncorrected) sample statistic

$$\frac{n(ac - bd)^2}{(a + b)(c + d)(a + d)(b + c)}$$

and referring it to the $\chi^2$ table with one degree of freedom. Note that both the numerator and the denominator differ from those of the conventional $\chi^2$ test discussed in an earlier section. Rejecting the hypothesis of relative symmetry implies that the index $[P_{01}/P_{0.} - P_{10}/P_{1.}]$ is nonzero; the index may, of course, be either positive or negative. Later, we propose the use of this index in "before and after" studies of item sensitivity and comment on its interpretation.

*Some Conventional Indices of Item Agreement and/or Association.* A major point we have made in the discussion of significance tests that might be applied to the various displays that we have considered is that, although the test of the formal hypothesis may be informative in a given situation, we often need in addition some measure of agreement and/or association which describes direction and strength. In other words, a conclusion of lack of independence, of lack of relative symmetry, or the like, which the test yields is not sufficient for certain purposes. In this section, we describe several generally well-known indices of this type for two-by-two tables.

We list below fifteen selected indices; all of them are expressed in terms of observed (sample) frequencies. The first is not a measure of association or agreement, but focuses instead on the marginals of the two items or administrations. It is the index of departure from equal sample difficulties in a display like Display A or Display B and is simply $(c - b)/n$. When we are interested in the relative difficulties, generally we would first run the McNemar test to determine whether or not the observed difference $(c - b)/n$ should be regarded as evidence of different difficulties in the population. If the test is not significant for some acceptable level of power, then $(c - b)/n$ should be regarded as merely a random departure from zero. The next five are measures of agreement or conditional agreement, the next six are measures of association in one sense or another, and the last three are nonsymmetric measures which arise when one deliberately focuses on columns in relation to rows or rows in relation to columns.

1. Index of departure from equal difficulties:

$$\frac{a + c}{n} - \frac{a + b}{n} = \frac{c - b}{n}$$

2. Proportion of agreement:

$$\frac{a + d}{n}$$

3. Ratio of observed proportion of agreement to the maximum proportion of agreement:

$$\text{Larger of} \quad \frac{a + d}{a + d + 2b} \quad \text{or} \quad \frac{a + d}{a + d + 2c}$$

4. Kappa: This is a "corrected for chance" proportion of agreement. See Cohen, (1960).

$$\frac{2(ad - bd)}{(a + b)(b + d) + (a + c)(c + d)}$$

5. Scott's coefficient: This also is a corrected proportion of agreement using an expected value that differs from the one used in Kappa. See Scott, (1955).

$$\frac{2ad - \frac{1}{2}(b + c)^2}{2ad + \frac{1}{2}(b + c)^2 + (a + d)(b + c)}$$

6. Ratio of observed covariance to maximum covariance: This also is the ratio of Phi to Phi Max or of Kappa to Kappa Max. See Horst, (1966), pp. 238-239; Berry *et al*, (1974).

$$\text{Larger of} \quad \frac{ad - bc}{(a + b)(b + d)} \quad \text{or} \quad \frac{ad - bc}{(a + c)(c + d)}$$

These two terms are two measures of conditional agreement described by Bishop *et al*, (1975), pp. 397-398. Note that Kappa is given by the ratio of the sum of the numerators to the sum of the denominators.

7. Phi:

$$\frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} = \sqrt{\frac{\chi^2}{n}}$$

8. Proportion of explained variance: See Bishop *et al*, (1975), pp. 389-390, Goodman and Kruskal, (1954), pp. 759-760. For a two-by-two table this is the same as Phi squared, i.e.,

$$\frac{(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

9. Symmetric Lamda: See Goodman and Kruskal, (1954), pp. 742-745 and 757-758. The formula depends upon the relative values of $a$, $b$, $c$, and $d$. As a special case, when $a > b < d$, $a > c < d$, and $a > d$, we have Symmetric Lamda =

$$\frac{2d - b - c}{2d + b + c}$$

This is the same as their measure of reliability in the unordered case, pp. 757-758.

10. Odds ratio: See Fleiss, (1973), pp. 43-49; Mantel and Hankey, (1975). A dual exists for the odds, but not for the odds ratio.

$$\frac{ad}{bc}$$

11. Yule's Q: See Fleiss, (1973) p. 45. This is a function of the odds ratio. It also is Gamma (Goodman and Kruskal, 1954, p. 750) for the case of a two-by-two table.

$$\frac{ad - bc}{ad + bc}$$

12. Squared Tryon Coefficient of Domain Validity: See Kaiser and Michael, (1975).

$$\frac{4(ad - bc)}{(a + d)(b + c) + 4ad}$$

13. Peirce's Theta: See Goodman and Kruskal, (1959), pp. 129-130. This is the difference between sample values for two conditional probabilities. A dual exists, since Theta is not a symmetric measure.

$$\frac{a}{a + c} - \frac{b}{b + d} = \frac{ad - bc}{(a + c)(b + d)}$$

$$\frac{a}{a + b} - \frac{c}{c + d} = \frac{ad - bc}{(a + b)(c + d)}$$

14. Lamda: See Goodman and Kruskal, (1954), pp. 740-742. The formula depends upon the relative values of *a*, *b*, *c*, and *d*, and on whether the prediction is for columns or for rows; thus a dual exists. Twelve different formulas exist; for four of these, Lamda has a value of zero. One formula is:

$$\frac{a - b}{a + c}$$

15. Index of departure from relative symmetry: This is the difference between sample values for two conditional probabilities, but is not identical with Peirce's Theta. Again, a dual exists. See Bishop *et al*, (1975), Chapter 7.

$$\frac{a}{a + b} - \frac{d}{c + d} = \frac{c}{c + d} - \frac{b}{a + b} = \frac{ac - bd}{(a + b)(c + d)}$$

$$\frac{a}{a + c} - \frac{d}{b + d} = \frac{c}{a + c} - \frac{b}{b + d} = \frac{ab - cd}{(a + c)(b + d)}$$

Certain relations among these indices are made evident by expressing each one in terms of the sample values of the two-by-two table. For example, seven of them (numbers 4, 6, 7, 8, 11, 12, 13) have as a numerator the quantity ($ad - bc$) which is the determinant of the matrix of observations given in Displays A through C; this determinant is zero (and thus the index is zero) whenever independence holds for the matrix. This notation raises the question of how different from each other these seven indices which are functions of this determinant actually are. Apparently, they differ primarily in a scaling factor which determines the largest and the smallest possible value for the index. The numerators of the dual values for the Index of Departure from Relative Symmetry, No. 15, also are determinants, but of rearranged matrices;

one of these appears in Display D (Observations) and Display E (Observations).

Another well-known function of a matrix like that of Display A (Observations) is the trace, or sum of the diagonal elements; this is $(a + d)$ which appears as the numerator in two of the indices. One is the proportion of agreement, No. 2, and the other is this same proportion divided by the maximum proportion of agreement which the marginals permit, No. 3. This same principle of adjusting the observed index in terms of the marginals appears in the Ratio of Observed Covariance to Maximum Covariance, No. 6. Perhaps the most interesting feature of this ratio is that it is the adjusted value of both Kappa, No. 4, and Phi, No. 7, thus, giving a link between a measure of agreement (Kappa) and a measure of association (Phi). Phi is a chi-square-like index which is isomorphic with the Pearson coefficient of correlation for two binary variables; as such it is a measure of association in a non-directional sense. Note that Phi-squared also appears as an index.

It is rather important to distinguish agreement from association. In particular, for a three-by-three table, with categories for both dimensions similarly ordered, agreement would generally be regarded as a function of the frequencies in the diagonal cells. A marked degree of association could exist in such a table even though these diagonal cells were empty. As another example, in a two-by-two table we have a high degree of association but no agreement when all the frequencies fall in the off-diagonal cells. If we are interested in a measure of agreement for the two-by-two table, we may choose among the proportion of agreement or this proportion divided by its maximum, a corrected for chance measure of agreement using either Kappa or Scott's coefficient, or two measures of conditional agreement. We have already mentioned that Kappa divided by maximum Kappa, for a two-by-two table, is identical to Phi/Phi Max.

Yule's Q, which is equivalent to the Goodman and Kruskal Gamma for two-by-two tables, is a function of certain conditional probabilities. For Display B (Parameters) the probability of being correct on the second administration, given a correct response on the first, is $P_{11}/P_{1\cdot}$, and the probability of being incorrect on the second administration, given a correct response on the first, is $P_{10}/P_{1\cdot}$. The ratio of these two values is $P_{11}/P_{10}$, and is called the odds for a correct response on the second administration when the item is answered correctly on the first administration. An analogous term $P_{01}/P_{00}$ is the odds for a correct response on the second administration when the item is answered incorrectly on the first administration. The odds ratio then is given by $P_{11}P_{00}/P_{10}P_{01}$ and is estimated by $ad/bc$. The odds ratio is a symmetrical index, but the odds are directional. Yule's Q is simply the difference between the two odds divided by their sum and is

$$\frac{P_{11}/P_{10} - P_{01}/P_{00}}{P_{11}/P_{10} + P_{01}/P_{00}} = \frac{P_{11}P_{00} - P_{01}P_{10}}{P_{11}P_{00} + P_{01}P_{10}}.$$

The sample analog is $(ad - bc)/(ad + bc)$.

The squared Tryon coefficient of domain validity, like several other indices, is a function of $(ad - bc)$ or the determinant of the entries in Display A (Observations). It is equivalent to KR-20 and Cronbach's Alpha for the special case of two binary items, and thus can be considered as an index (strictly, a lower bound) of "reliability." It is simply the squared Pearson product-moment correlation between a variable defined as the sum of scores on the two items and a variable defined as the sum of scores on all items in the domain (including those two). This latter is a hypothetical variable; the assumptions needed to derive the formula under these conditions are discussed by Kaiser and Michael (1975). Because the squared Tryon index is a measure of relation between two items and the complete domain of items, it seems of little use as a measure either of agreement or association for only the two items themselves.

The non-symmetric indices are appropriate when prediction (or conditional probability) is directional. For example, Peirce's Theta is a function of a prediction model (e.g., predicting response on the second administration from the response on the first) which assumes that the observed frequencies arise from a mixture of perfect prediction and the operation of an extraneous chance device that predicts 1 with a specified probability and predicts 0 with the complement of that probability. The two values of Theta correspond to the two directions of prediction (e.g., columns from rows or rows from columns); both have the same numerator. Two weights, which sum to unity, are assumed to determine this mixture of perfect prediction and chance. Theta is the mixture weight for the perfect prediction component, and it also is a difference between sample values of two conditional probabilities. Theta is a measure of association in the sense of prediction success.

Lambda and Symmetric Lambda are measures of predictive association. Lambda is a directional measure which gives the proportional reduction in the probability of error in predicting the column (row) category, knowing the row (column) category. Symmetric Lambda is a function of the numerators and denominators of the two directional indices for the given sample frequencies; it gives the proportional reduction in the probability of error in predicting either category, knowing the other. Symmetric Lambda always takes a value between the values of the two directional Lambdas.

We commented above on one of the two indices of departure from relative symmetry, pointing out that it may be interpreted as a measure of "learning" or "forgetting" for the study of performance on the same item before and after instruction (Display E). The index to use is

$$\frac{ac - bd}{(a + b)(c + d)}.$$

For this Display, relative symmetry implies that "learning" and "forgetting" offset each other; when

$$\frac{ac - bd}{(a + b)(c + d)}$$

is positive, there has been more "learning" than "forgetting." For a study of two different items there are two different measures of departure from relative symmetry that indicate slightly different relationships. The expression

$$\frac{ac - bd}{(a + b)(c + d)}$$

is conditional on the responses to Item A, and

$$\frac{ab - cd}{(a + c)(b + d)}$$

is conditional on the responses to Item B. We observe that these two indices are equal when $b = c$, i.e., when we have exact symmetry (equal difficulties) in the sample; they are equal in magnitude but opposite in sign when $a = d$. (If both $b = c$ and $a = d$, both indices are zero.) Since the two indices may be opposite in sign, averaging them to secure a composite index for the two items seems inadvisable.

We now wish to examine these fifteen indices from two points of view. First, we shall ask whether or not the index is independent of or adjusted in some fashion for the level of difficulty. Second, we shall consider the index as an estimator and in particular identify those that are, or can be modified to be, unbiased estimators.

Let us consider a situation in which the observed frequencies, instead of being $a$, $b$, $c$, and $d$ are systematic multiples of these numbers such as $awy$, $bwz$, $cxy$ and $dxz$; in other words, the entries in each of the original rows and columns have been multiplied by a (different) constant. From one point of view, the table of frequencies

| | |
|---|---|
| $awy$ | $bwz$ |
| $cxy$ | $dxz$ |

exhibits the same degree or amount of association as does the original set of frequencies

| | |
|---|---|
| *a* | *b* |
| *c* | *d* |

but is characterized by altered marginals. Indices that take the same value for these two tables are independent of this multiplicative alteration of the marginals, and are, in this sense, independent of item difficulties. It can readily be shown that none of the five measures of agreement have this independence, and that of the six symmetric measures of association, only the Odds Ratio and Yule's Q have it. In contrast, if we alter the marginals by multiplying only the rows (or only the columns) by different constants, one of the values of Theta and one of the indices of Departure from Relative Symmetry are unaffected.

A different approach is to adjust the index in some fashion in relation to marginals, or item difficulties. Four types of adjustment are represented in our collection of indices. Measures may be simple functions of conditional probabilities, i.e., the index may be a sum or difference of cell values that are conditional on marginal values, as are (13) and (15). Note that both of these are directional and exist as duals, depending on whether the row or the column marginals are adjusted for. Also note that the conditional probabilities involved in the difference are independent of each other. Measures may also be functions of ratios of conditional probabilities; for example, the Odds is a ratio of two conditional probabilities and the Odds Ratio (10) a ratio of two different ratios of conditional probabilities. Also, Yule's Q (11) is a function of the Odds Ratio and consequently of a ratio of two different ratios of conditional probabilities. Second, measures may be adjusted by relating the measure to the maximum value permitted by the marginals. This type of adjustment is represented by (3), which is the adjusted proportion of agreement, and by (6) which is both Phi/Phi Max and Kappa/Kappa Max. Third, an adjustment may be made by subtracting an expected value; Kappa (4) for example, is a normed difference between the observed proportion of agreement and the expected proportion of agreement given that the two classification procedures are independent and can differ in their population marginal proportions. Scott's coefficient (5) is similar, but computes the expected proportion of agreement on the assumption that the independent classification procedures do not differ from each other in their population marginal proportions. For both these, the norming divisor is 1 minus the expected proportion of agreement. Fourth, in Lambda (14) and in Symmetric Lambda (9) marginals play a role analogous to the expected values in (5), (6), with the marginals (proportions, not frequencies) defining the extent of prediction given no information about the result of the other classification. The two Lambda measures are oriented toward prediction, rather than measuring agreement, and are measures of association in this sense. We merely point out that indices (1), (2), (7), (8), and (12) have no correction or adjustment for the marginals.

We now consider each of the indices as an estimator and ask whether or not it yields an unbiased estimate of the analogous population parameter. Certain generalizations about expected values are useful here. For example, a sample proportion is known to be an unbiased estimator of the population proportion, and the expected value of a sum (or difference) of random variables is known to equal the sum (or difference) of the expected values of the random variables; this indicates that an index like (1) or (2) is an unbiased estimator. To illustrate, we give in Table 2 a population of eight paired responses and then for several biased estimators give numerical values for the mean of the sampling distribution (based on samples of size 7 without replacement) of the index, compared with the population value of the index. Note

**Table 2: A Sampling Experiment with a Finite Population**

| Responses: | Item A | Item B |
|---|---|---|
| | 1 | 1 |
| | 1 | 1 |
| | 1 | 1 |
| | 1 | 0 |
| | 1 | 0 |
| | 0 | 0 |
| | 0 | 0 |
| | 0 | 1 |

| *Index* | *Mean Value for Samples of Size 7* | *Population Parameter* |
|---|---|---|
| Larger of: $\frac{a + d}{a + d + 2b}$ or (3) $\frac{a + d}{a + d + 2c}$ | .7143 | .7201 |
| Kappa (4) | .2244 | .2500 |
| Phi/Phi Max (6) | .3958 | .3333 |
| Yule's Q (11) | .4786 | .5000 |
| Squared Tryon Coefficient (12) | .3781 | .4103 |

that Kappa is a biased estimator (in the non-null case), even though the expected value of Kappa is 0 when independence holds; Phi, of course, is well-known to be a biased estimator in the non-null case, as is any Pearson product-moment correlation coefficient. Also note that "correcting" Phi (or Kappa) in relation to its maximum value does not remove the bias.

For the Odds Ratio, the expected value of the sample ratio for samples of a fixed size is not necessarily equal to the parameter. (It also is true that the sample ratio will be indeterminate when $b$ or $c = 0$; this occurs once in our sampling experiment.) However, Mantel and Hankey (1975) show that $E(ad)/E(bc)$ and $E(a)E(d)/E(b)E(c)$ both equal the parameter for the case of $n$ fixed and for the case of fixed marginals for either (but not both) rows or columns. A similar generalization holds for Yule's Q; i.e.,

$$\frac{E(ad) - E(bc)}{E(ad) + E(bc)}$$

will equal the parameter, but the expected value of $(ad - bc)/(ad + bc)$ will not; we give these two numerical values for our sampling experiment in Table 2. Apparently we can secure an unbiased estimate of the population Q by splitting a sample of paired observations into two random samples (of equal or approximately equal size) and then averaging the $ad$ and $bc$ sample values. These averaged values would then be used to estimate the population Q. This "unbiasing" procedure is useful here when we deal with products; splitting samples and using the averaging values for estimation is unnecessary when the sample index itself is unbiased. Like many ratios, the squared Tryon Coefficient is biased.

We now wish to bring together the displays, the significance tests, and the indices we have discussed and use them to describe studies of item equivalence, item stability, and item sensitivity to instruction.

*Item Equivalence Studies.* Display A (Observations) and its modification Display D (Observations) describe arrangements appropriate for the study of responses to two different items by the same sample of students. If the items are drawn from the same item domain, the expectation would be that the items tend to measure "the same thing" and thus, that the hypothesis of independence is false. If so, the conventional $\chi^2$ test of independence could confirm this; however, the test does not yield information that would be given by a measure of agreement or association. Different items may differ in difficulty whether they are or are not characterized by marked agreement or association. If equal difficulty is regarded as all or a critical part of equivalence, then the McNemar test is required. However, if equivalence is viewed primarily as agreement or association, independent of difficulty, then the McNemar test is of minimal use. There are two tests of relative symmetry for the data of Display A; one is conditional on the responses to Item A and the other is conditional on the responses to Item B. It is rather attractive to regard relative symmetry as a type of equivalence for two items, since relative symmetry holds whenever the conditional probability of answering the two items the same way is the same for those responding correctly and those responding incorrectly to one of the items. However, these two tests can be inconsistent, just as the two measures of departure from relative symmetry can be opposite in sign, and so we would not recommend their use in this situation. Later, we will show that the logically correct test can be informative in a different type of study—that of sensitivity to instruction.

We now turn to indices of agreement and/or association. If items are drawn from the same item domain, we may regard the items as characterized by conceptual homogeneity (See Harris, 1974, p. 100-103). The question of their response homogeneity may also be of interest; in particular, we may wish to estimate a response homogeneity parameter for the domain as a whole as a descriptor of the domain's functioning at a specified point in the instructional process. Such a description or characterization could provide useful information about the teaching-testing complex. For example, response homogeneity for the domain might be estimated to be low early in the instruction and high later in the instruction; this evidence would be consistent with an expectation of consolidation or integration of the tasks that are taken as the evidence of achievement. If, however, the estimate remains at a low level throughout the instruction, then we have evidence that the instruction does not have this effect. As a second example, two somewhat different instructional programs, when both are designed to achieve the same objectives, might be associated with noticeably different estimates of response homogeneity for the same item domain; this evidence could be useful in characterizing the two instructional procedures. As a third example, two different item formats might reasonably be regarded as appropriate in generating an item domain for an instructional program; estimates of response homogeneity for the two types of domain would provide important data about the equivalence (or lack of equivalence) of the two formats.

The problem we face is essentially this: For the various pairs of items in the population of items we can conceptualize a symmetric matrix, with the off-diagonal elements of this matrix containing the values of the index for the various item pairs. (What is called multiple matrix sampling can be employed; this procedure does not necessarily use the same sample of individuals to estimate the index value for any two item *pairs*.) Given satisfactory estimates of each index value, the problem is then one of deriving a secondary index that characterizes all these values, probably estimating it by employing only a random sample of the item pairs. It seems natural to consider an average of these indices of item pairs as the simplest secondary index, and we shall proceed on this assumption. Further work on this problem of a secondary index seems necessary; we hope to address it some time in the future.

Among the measures of agreement, the proportion of agreement, index (2), is an estimate of a parameter that when averaged over all item pairs yields a readily interpretable secondary index. The fact that index (2), for a single item pair, is an unbiased estimator also recommends it. The fact that the index is not independent of the item difficulties may argue against its use, but this argument can be countered in the following way.

In any study of equivalence of a set of item pairs, with this set a proper sample of the item pairs making up the domain, the administration is likely to occur at a particular point in the instructional process. To a considerable extent, the choice of this point in time determines the difficulty of the items studied; thus, early in the instruction it is likely that many of the items are (relatively) difficult. (This should not be taken to imply that the several items are equal in difficulty, though if they are, the statement still holds.) Insofar as the difficulty of the members of an item pair imposes a limit on the percent of agreement, one may expect a smaller percent of agreement for a pair early in the instruction than later; if so, a secondary index taken as an average proportion of agreement would be specific to the point in the instructional process when the study was made. This is not undesirable, however, since the study also yields unbiased difficulty estimates that can be reported and interpreted along with the proportion of agreement. Thus, we could find the average proportion of agreement low and average difficulty high early in the instruction and average proportion of agreement high and average difficulty low later; this would be evidence of a particular pattern of learning. If we use the adjusted proportion of agreement, these same data might indicate relatively little change in the proportion of agreement in relation to its maximum; further, this adjusted proportion of agreement is not an unbiased estimator. This line of reasoning leads us to regard the "raw" proportion of agreement, index (2), as a useful measure of equivalence. We would prefer it to Kappa or Scott's Coefficient on similar grounds.

Of the measures of association, Yule's Q, No. 11, probably is the most useful. We prefer it to the Odds Ratio on the grounds that the sample Odds Ratio is indeterminate whenever *b* or *c* is zero. We interpret Yule's Q as the Goodman-Kruskal Gamma, which it equals in the two-by-two case. They say that Gamma is

> The *difference* between the conditional probabilities of like and unlike orders . . . when two individuals are chosen at random from the population. (Goodman and Kruskal, 1954, p. 749.)

As such, Gamma tells us how much more probable it is to get like than unlike orders in the two classifications (items). It seems meaningful to average this difference in conditional probabilities over item pairs to secure a secondary index for the item domain. We commented above on the possibility of securing unbiased estimates by splitting the sample. Phi or functions of Phi, like (6) and (8) or the correlation coefficient of (12), probably should not be averaged in this type of situation. Note that one might take the median of the observed product-moment correlation between two specified variables for several samples of individuals as an estimate of the correlation for the population. This in no way implies that correlations between different pairs of variables, either based on the same sample or on different samples, yield a distribution for which a measure of central tendency has a meaningful interpretation. Symmetric Lambda gives the proportionate reduction in the error of prediction when we know either classification (item result), as compared with knowing only the marginals. As such it is adjusted for the marginals; however, it is not an unbiased estimator. In addition, averaging values of Symmetric Lambda for item pairs may be difficult to defend, just as averaging product-moment correlations for different pairs of variables is difficult to defend, especially in the absence of rather strong structural assumptions about the items. For these reasons, we would not recommend the use of Symmetric Lambda over Yule's Q.

These comments indicated a case that can be made for the proportion of agreement and for Yule's Q as measures of equivalence of items; we emphasize, however, that one is a measure of agreement and the other a measure of association and as such represent different aspects of equivalence (which is a term of many meanings). A choice between these two aspects still must be made if one is to use only a single secondary index to characterize the item domain; possibly the best strategy (since it would add little cost) would be to develop both indices and report them. There is a third type of index that differs somewhat from agreement or association as represented by the indices we have discussed here. This third type rests on a latent structure conception and provides estimates of the probability of an inappropriate response (answering correctly when he does not know and answering incorrectly when he does) for each item. This third type of index and its estimation is described in considerable detail in Chapter 5.

*Item Stability Studies.* Display B (Observations) and Display E (Observations) describe arrangements appropriate for the study of responses to the same item on two occasions closely related in time during which essentially no learning is assumed to take place. In this situation, the hypothesis of independence of the responses to the same item on two occasions is unlikely to be true, and so running the conventional $\chi^2$ test usually would be an academic exercise.

In an item stability study, we would expect that the item difficulty parameters would be identical on the two occasions; this is the assumption of essentially no learning during the interval between administrations. It also is the hypothesis of exact symmetry. The McNemar test is designed to test this hypothesis and thus, might be used here to identify items for which item difficulty is altered over the two administrations. (See our earlier discussion of the McNemar test.) If item difficulty is altered, an explanation should be sought. For example, this might result from the use of an unfamiliar item type of format (without adequate pre-test instruction) which might depress scores artificially on the first administration. Another example might be the occurrence of a fortuitous event during the interval that gives the student a new view or understanding of the content of the item. We suggest that when the hypothesis of exact symmetry is rejected for a reasonable sample size, no further analyses of these data be made. Instead, an attempt should be made to develop a plausible explanation of this result in terms of item-type characteristics or requirements and/or instructional procedures. Such an explanation should lead to a modification of the item generation principles employed and/or of the instructional program. The same or the modified item can then be studied anew later; the result of such a study should support or discredit the explanation that was developed. We recognize that it is possible that idiosyncratic items which shift in difficulty for no discernable reason may exist; however, deciding that an item is of this type should be a conclusion of last resort.

For a measure of agreement in an item stability study, we would recommend the sample proportion of agreement, index (2), for the same reasons we recommended it for equivalence studies. It is an unbiased estimator, it may be averaged meaningfully over items, and the item difficulty level—though limiting the value of the index—can be estimated and reported separately to provide a notion of the relation of the proportion of agreement to its maximum. For a measure of association, we also recommend Yule's Q, with the interpretation of Gamma given above, for the reasons we discussed in the previous section. Again, both indices might be employed in any study. There also is a third type of index, developed in terms of latent structure, that provides an estimate of the probability of an inappropriate response. This third type is discussed extensively in Chapter 4.

*Studies of Sensitivity to Instruction.* Two points will help describe the position we take with respect to studies of sensitivity to instruction. First, we believe that it is reasonable to expect that student responses to the item universe that is developed in connection with an instructional program will be altered as a function of the instruction. Second, we believe that a characterization of the item universe (or a sub-universe of interest) with respect to sensitivity to instruction is an important datum for describing and possibly modifying the total teaching-testing process. We imply that this desired characterization of the item universe may be estimated by studying only a sample of items and then developing from the data for these items an appropriate index. We do not imply, however, that items which lack sensitivity to instruction are to be discarded; in other words, we do not propose that sensitivity to instruction become an item selection technique. This is consistent with our desire to find ways to study the teaching-testing complex that will result in an improvement of the total complex itself.

The expectation that student responses to the item universe will be altered as a function of the teaching can be tested empirically. There are two rather obvious designs we might employ. The first is a prepost design. The second is a comparative prospective study. In the prepost design, the item is given to a sample of students before they have received the relevant instruction and again following instruction. The relevant displays are Display B and Display E, with the understanding that the two administrations are separated by instruction, which may occupy considerable time and thus separate the two administrations by a substantial period of time. For the prepost design, we select a sample of $n$ students who have not yet been instructed with the particular program, but these should be students for whom the program is appropriate. The item is then administered before and after instruction. In this design, $n$, the sample size, is at our choice, but the marginals are determined by the results of the experiment, with $(a + b)/n$ estimating the difficulty ("easiness") of the item prior to instruction and $(a + c)/n$ the difficulty following instruction. If the instruction has moved the students in the direction of competence with respect to this item, then we expect that $(a + c)$ will be greater than $(a + b)$. The McNemar test is designed to test the hypothesis of equal pre and post difficulties in the population in this type of experiment, and may be used here; generally, we would expect to reject this hypothesis. However, note that the probability value that one secures from the test is not a metric that describes the degree of change; an additional statistic is needed.

We merely state that neither the conventional chi-square test of independence nor the related phi coefficient supplies such a statistic. Two observations are relevant. Both the chi-square test and the phi coefficient are symmetric functions and thus ignore the fact that responses prior to instruction are necessarily antecedent, and those following the introduction of the manipulated experimental variable that is labeled "instruction" are necessarily subsequent. "Before" and "after" cannot be ignored in this design. Second, insofar as chi-square and phi are satisfactory measures of association, they are inappropriate here because we wish to measure a change in response rather than a stability of response (lack of independence) which would seem to be the obvious interpretation for a large value of chi-square or phi. We also merely mention that neither the observed proportion of agreements, given by $(a + d)/n$, nor a "corrected" proportion of agreements such as that given by Kappa is informative here; agreement can be distinguished from association, but neither provides the measure we seek.

The simplest measure of "learning" and "forgetting" would be $(c - b)/n$, which is the difference between the sample proportion who learn and the sample proportion who forget. This is the Index of Departure from Equal Difficulties, or (1) in our list. We may "conditionalize" a measure like this in the following fashion. The observed probability of responding correctly, given that the student responded incorrectly prior to instruction, is given by $c/(c + d)$, which is a binomial variable that estimates the population value of the conditional probability of changing the response in the desired direction. A second sample value is $b/(a + b)$ which estimates for the population the probability of responding incorrectly following instruction, given that the student responded correctly prior to the instruction; this is the "negative" measure describing the probability of changing the response in the "wrong" direction. Both these sample values are unbiased and maximum likelihood estimates of the population parameters. These two sample values may be combined into a single index, thus: $c/(c + d) - b/(a + b)$, which is the same as

$$\frac{ac - bd}{(c + d)(a + b)}.$$

We observe that this is simply the logically correct Index of Departure from Relative Symmetry, No. 15, for this situation. We might, if we wished, first run the $\chi^2$ test on Display E (Observations) in order to test the hypothesis of relative symmetry; this is the hypothesis that "learning" and "forgetting" offset each other. Generally, we would expect this hypothesis to be false, and thus, a significant test result would not be unexpected; given such a result, we would then regard the observed value of

$$\frac{ac - bd}{(a + b)(c + d)}$$

as a nonrandom departure from zero. The observed value of

$$\frac{ac - bd}{(a + b)(c + d)}$$

estimates the difference in conditional probabilities $(P_{01}/P_{0.} - P_{10}/P_{1.})$

for the population; these conditional probabilities are independent of how easy or how difficult the item was at the time of the pre-test. We now propose an additional step, that of averaging such estimates over a random sample of *items* to secure a secondary index that describes the functioning of the item domain in this instructional program. This average over items should be an estimate of the average of the population values.

One might also treat the data of the "before and after" study in the manner described by Marks and Noll (1967). Their model assumes that when there is no "error" in responding, the responses to the item before and after instruction form a Guttman scale with only three response patterns appearing. This is equivalent to their assumption that the student does not "unlearn" the response. Marks and Noll show how the number of persons fitting each of the three admissable patterns can be estimated from the data. We could then use these "corrected" frequencies as a basis for estimating the conditional probability of changing the response in the "right" direction. (The other conditional probability is zero for this model.) With regard to the observed values, this "corrected" conditional probability estimate is $(c - b)/(c + d)$. One problem with the Marks and Noll model is, as they show, that it is possible to secure actual item data for which "corrected" frequencies turn out to be negative and thus, are inadmissible.

A second design for the study of sensitivity to instruction selects a sample of students who have received the instruction and a sample who have not. The item is then administered to the two groups and the data summarized as in Display C. This is a prospective study, in Fleiss's terms (1973, pp. 15-19), and differs from the first design which he calls a cross-sectional study. An important caution is that the instructed and uninstructed samples should be drawn from populations that are alike in all possible respects except for the presence or absence of the instructional experience. In practice, this may be very difficult to accomplish. For example, if the instructed group were drawn from tenth grade students who were completing a year's work in algebra and the uninstructed group drawn from students beginning tenth grade, we would have an obvious systematic difference in age or maturity in the populations which might influence the findings. As another example, if the same instructed group were compared with students drawn from those completing the tenth grade who had not taken algebra, we could have systematic differences in preferences for school subjects and possibly in academic ability between the two populations. Again, these systematic differences might influence the outcome of the study and obscure the relation between presence or absence of instruction and item performance. In our view, these sampling problems make the pre and post design a preferred one.

In the prospective design, the marginal frequencies for the two samples ($a + b$ and $c + d$) are at our choice and this makes it desirable to select a test and/or an estimator that is not sensitive to these particular marginal values. However, the marginals for the item ($a + c$ and $b + d$) are functions of the overall effectiveness of the instruction, the difficulty of the item, etc., and we usually would not want to adjust for them. The Proportion of Agreement, $(a + d)/n$ is a "natural" index of the extent to which instruction is effective and non-instruction is not effective in teaching the proper response to the item. We can adjust this index for the sample sizes of the two groups (Instructed and Not Instructed) by using the sum of two conditional probabilities,

$$\frac{a}{a+b} + \frac{d}{c+d} = \frac{ac + 2ad + bd}{(a+b)(c+d)},$$

as the corrected proportion. Note that we do not adjust for the maximum proportion of agreement, as in Index (3), since the maximum is a function of both marginals. Also note that we here use a sum of these two conditional probabilities; their difference, which is an index of Departure from Relative Symmetry, No. 15, is not informative in this design. We also might use

$$\frac{a}{a+b} - \frac{c}{c+d} = \frac{ad - bc}{(a+b)(c+d)},$$

or Peirce's Theta, as an index; in this situation it describes the difference in the conditional probability of getting the item right for those who were instructed and those who were not instructed. As such, a large value of Theta indicates effective instruction. For this design, we might also use the odds ratio or Yule's Q which is a function of the odds ratio. For example, $a/b$ gives the odds of getting the item right for the Instructed sample, and $c/d$ the odds for the Not Instructed sample. The ratio of these two odds is $ad/bc$. A large value indicates greater odds of getting the item right for the Instructed sample. Also note that the odds ratio (and consequently Yule's Q) is independent of any arbitrary weighting of the marginals. The odds ratio has the further feature of being interpretable in terms of a logistic model. In setting up a prospective study of this type, it is a good policy to choose equal size samples of instructed and noninstructed students; this gives maximum precision of the estimated odds ratio (Fleiss, 1973, p. 55). We see no strong arguments for preferring one of these indices; empirical studies probably are needed.

Fleiss (1973, Chapter 6) also describes a retrospective study design which, in this instance, would study samples of students who answer the item correctly and students who answer it incorrectly to determine whether they were instructed or not. This type of study is well-known in biomedical research; Fleiss describes it as the identification of the two study samples on the basis of the presence or absence of the outcome factor (e.g., performance on the item), and the estimation for both samples of the proportions possessing the antecedent factor under study (e.g., instruction or its absence). Fleiss comments that a greater degree of ingenuity is needed for the proper design of a retrospective study than for the design of a prospective study. We would not recommend the retrospective study as a basis for estimating sensitivity of items to instruction.

## Chapter 4

# NEW METHODS FOR STUDYING STABILITY

**RAND R. WILCOX**

Our goal in this paper is to describe probability models that may be useful in stability studies of achievement test items. The first of the three models we are about to describe has its origin in the work of Paul F. Lazarsfeld and Patricia Kendall as described by Goodman and Kruskal (1959, pp. 149–152). The model described by Lazarsfeld and Kendall would fit any situation in which people are asked the same yes-or-no question at two different times. In this setting, it is supposed that there are really two classes of people in the population of interest, those who have a tendency to answer "yes," in proportion $k$, and those who have a tendency to answer "no," in proportion $(1 - k)$. It is further supposed that the answers people give do not always represent their "true" tendency since they may be temporarily swayed in the other direction, may misunderstand, etc. Suppose that the "yes" people answer "no" with probability $x$, and that the "no" people answer "yes" with probability $y$. It is further assumed that both $x$ and $y$ are less than or equal to ½.* If we choose at random a member of the population, then the probabilities of the four possible outcomes are:

| | Second Answer | |
|---|---|---|
| First Answer | $P_{11} = k(1 - x)^2 + (1 - k)y^2$ | $P_{10} = k(1 - x)x + (1 - k)y(1 - y)$ |
| | $P_{01} = k(1 - x)x + (1 - k)(1 - y)y$ | $P_{00} = kx^2 + (1 - k) \cdot (1 - y)^2$ |

For this model, there is no unique solution for $x$, $y$ and $k$ in terms of $P_{ij}$ $(i, j = 0, 1)$, since we have in effect two equations with three unknowns. A modification consisting of assuming that $x = y$, gives the Lazarsfeld-Kendall "turnover" model for which a solution (in terms of parameters) exists.

We now adapt the solvable Lazarsfeld-Kendall (henceforth, L-K) turnover model to a specific achievement test item and a population of examinees. The following assumptions are made:

1. A total of $n$ examinees is randomly selected from the population of examinees and each of the selected examinees answers the test item on two different occasions which are not widely separated in time.

---

*A table of symbols appears at the end of this chapter.

2. The sampling of examinees is from an infinite population or from a finite population with replacement.
3. There are two classes of people in the population of examinees, namely, those who know the answer to the test item and those who do not. Furthermore, each of the $n$ examinees selected either knows the answer on both of the occasions or does not know the answer on either of the occasions that he is tested. We deal with this assumption at least partially by proposing that the McNemar test of exact symmetry be applied to sample data, and the L-K model be employed only if this test is not significant. We also provide a table of sample sizes for which the power of the McNemar test is .50 for various alternatives to exact symmetry; this table can be used to aid in interpreting the results of the McNemar test.
4. A person who knows the answer to the test item responds incorrectly with probability $x \leq 1/2$ and a person who does not know the answer to a test item responds correctly with probability $x \leq 1/2$.
5. Given that an examinee knows the answer to the test item, an incorrect response on the first occasion is independent of an incorrect response when retested. Correspondingly, for an examinee who does not know the answer, the event of a correct response on the first occasion is independent of the event of a correct response on the second occasion.

We denote a correct response with a 1 and an incorrect response with a 0. Let $k$ denote the proportion of people in the population of examinees who know the answer to the test item. The probabilities corresponding to the four possible outcomes (right-right, right-wrong, wrong-right, and wrong-wrong) are then:

$$\begin{aligned} P_{11} &= x^2 - 2kx + k \\ P_{10} &= x(1 - x) \\ P_{01} &= x(1 - x) \qquad\qquad (1) \\ P_{00} &= x^2 - 2x(1 - k) + (1 - k) \end{aligned}$$

$$P_{11} + P_{10} + P_{01} + P_{00} = 1$$

Note that $P_{11}$, $P_{10}$, $P_{01}$, $P_{00}$, $x$, $y$, and $k$ are population parameters.

What we observe is a fourfold table of frequencies:

| | 1 | 0 | |
|---|---|---|---|
| 1 | $a$ | $b$ | $a + b$ |
| 0 | $c$ | $d$ | $c + d$ |
| | $a + c$ | $b + d$ | $n$ |

where $a + b + c + d = n$. Note that $P_{10} = P_{01}$ but it is not necessarily true that $b = c$. With the parameters $P_{10}$ and $P_{01}$ equal, we have the case of exact symmetry, or equal difficulties for the item on the two occasions. Our first step is to use the McNemar test with the observed frequencies to test this hypothesis of exact symmetry. In general, we expect the normative difficulties of the item to be the same on the two occasions unless some unintended learning has taken place in the relatively short interval. It should be clear that the power of the McNemar test is a critical factor in its use and interpretation. It is well known that if a stated hypothesis is false, then, given a large enough sample, it will generally be possible to reject this hypothesis at a chosen significance level; however, we wish to guard against rejecting this hypothesis when the normative item difficulty is only trivially different on the two occasions. We have used results of Bennett and Underwood (1970), who discuss the power function of $X^2 = (b - c)^2/(b + c)$ under alternatives of the form $P_{10} = q + e/\sqrt{n}$, $P_{01} = q - e/\sqrt{n}$, where $e$ is some constant. The limiting power function is a non-central chi-square distribution with 1 *df* and non-centrality parameter $\lambda = (2e^2)/q$. In the limiting power function, $\lambda$ depends on $\beta$, the power of the test. See expression (9) in Bennett and Underwood (1970, p. 341). *If* $\alpha = .05$ and $\lambda = 3.84$, then $\beta = .5$, which can be read from the tables given by Fix (1949). Now $P_{10} - P_{01} = (2e)/\sqrt{n}$, and $e = [(\lambda q)/2]^{1/2}$. Therefore, for $\alpha = .05$ and $\beta = .50$, $n = [2/(P_{10} - P_{01})]^2(3.84/2)q$. It then follows that for the chosen $\alpha$ and $\beta$ we can table $n$ in terms of the parameters $q$ and $(P_{10} - P_{01})$. If we wish $\alpha$ to equal .01, $\lambda = 6.64$ and the new values of $n$ are readily given. Table 1 gives values of $n$ for various values of the parameters $q$ and $P_{10} - P_{01}$ that yield a probability of .5 of falsely rejecting the hypothesis of exact symmetry at the .05 level of significance.

**Table 1: Sample Sizes Needed for the Power of the McNemar Test To Be 0.5 when the Significance Level is 0.05.**

| | | $P_{10} - P_{01}$ | | |
|---|---|---|---|---|
| $q$ | .1 | .2 | .3 | .4 |
| .10 | 76 | 19 | * | * |
| .15 | 115 | 28 | 12 | * |
| .20 | 153 | 38 | 17 | 9 |
| .25 | 191 | 48 | 21 | 12 |
| .30 | 230 | 57 | 25 | 14 |
| .35 | 268 | 67 | 29 | 16 |
| .40 | 307 | 76 | 34 | 19 |
| .45 | 345 | 86 | 38 | 21 |

*Impossible, since $P_{01}$ would be negative.

We now examine the solvable Lazarsfeld-Kendall model. From (1) above we have $(x - x^2) = P_{10} = P_{01}$, and so, a solution for $x$ in terms of parameters is $\frac{1}{2}[1 \pm (1 - 4P_{10})^{1/2}]$. A solution for $k$ is

$$\frac{1}{2} + \frac{2P_{1\cdot} - 1}{2(1 - 4P_{10})^{1/2}}.$$

Note that $P_{1\cdot} = P_{11} + P_{10} = x^2 - 2kx + k + x(1 - x) = x - 2kx + k$. In terms of parameters, $2x(1 - x)$ is the probability that a randomly chosen person will answer the test item differently on the two separate occasions; this suggests a reason for Lazarsfeld and Kendall's choice of "turnover" as a description of the model.

The necessary and sufficient conditions for the L-K model to hold are:

$$\begin{aligned} &P_{01} = P_{10} \leqslant \tfrac{1}{4} \\ &P_{11} \geqslant P_{1\cdot} P_{\cdot 1} \end{aligned} \qquad (2)$$

For convenience, set $p = P_{10} = P_{01}$. We now develop $\hat{p} = (b + c)/(2n)$ as an estimator of $p$. Define the random variable $A$ as the number of examinees that answer the test item correctly on both occasions. Lower case letters corresponding to the symbol used for a random variable will designate a value of the random variable. Thus, $a$ is an observed value of the random variable $A$. Let $B$ be the random variable that corresponds to the number of students who answer correctly on the first occasion but incorrectly on the second occasion. The value of $B$ which we observe is $b$. $C$ and $D$ are random variables that are defined in a similar fashion for students who answer incorrectly on the first occasion or on both occasions, respectively. As stated earlier, it is assumed that the subjects are sampled from an infinite population or from a finite population with replacement. Therefore the appropriate joint probability density function (*p.d.f.*) for $A$, $B$, $C$ and $D$ is the multinomial with parameters $n$ and $P_{ij}$, $i = 0, 1$ and $j = 0, 1$. That is, the probability that $A = a$, $B = b$, $C = c$ and $D = d$ is given by

$$\frac{n! P_{11}^{a} P_{10}^{b} P_{01}^{c} P_{00}^{d}}{a! b! c! d!}, \qquad (3)$$

where $a + b + c + d = n$ and $P_{11} + P_{10} + P_{01} + P_{00} = 1$. Since by assumption $p = P_{10} = P_{01}$, expression (3) is equivalent to

$$\frac{n! P_{11}^{a} p^{b+c} P_{00}^{d}}{a! b! c! d!}. \qquad (4)$$

Let $W$ designate the random variable which corresponds to the number of students from a sample of $n$ students who give inconsistent answers on the two occasions they are tested. That is, $W$ denotes the number of students who are correct on one of the two occasions and incorrect on the other. The probability of securing an inconsistent response from a randomly selected student is $P_{10} + P_{01} = 2p$. Given that sampling is from an infinite population or from a finite population with replacement, $W$ has a binomial distribution with parameters $2p$ and $n$. Note that the observed value of $W$ is $W = b + c$ and that the observed value of $W/(2n)$ is $\hat{p} = (b + c)/2n$. Since $W$ has a binomial distribution with parameters $n$ and $2p$, the expected value of $W$ is $2np$. Then, $E(W/2n) = 2np/2n = p$. Hence, $W/2n$, or equivalently, $\hat{p}$, is an unbiased estimate of $p$.

Next we can show that $W/2n$ is a maximum likelihood estimator (MLE) of $p$. A MLE is that estimator of the parameter which makes the observed values most likely. For example $a/n$ is well known to be a MLE of $P_{11}$, i.e., $a/n$ is that estimate of $P_{11}$ which makes the event $A = a$ most likely.

A theorem due to Zehna (1966, p. 744) shows that under very general conditions, when $\hat{\theta}$ is a maximum likelihood estimator of the parameter $\theta$ and $g$ is any (measurable) function, then $g(\hat{\theta})$ is a maximum likelihood estimator of $g(\theta)$. Zehna shows that the function need not even be one-to-one, which is a condition for which it was earlier proved that $g(\hat{\theta})$ is a MLE of $g(\theta)$. Let $(X, B)$ denote a measurable space (Loeve, 1963, p. 64) and let $T = \{P_\theta;\ \theta \in \Theta\}$ be a family of probability measures on $(X, B)$. The parameter space $\Theta$ is an interval in an $r$-dimensional Euclidean space $(r \geqslant 1)$. Thus, $\theta = (\theta_1, \ldots, \theta_r)$ is vector valued. The set $T$ could, for example, be the class of probability functions of the form $\theta^z(1 - \theta)^z$, $0 \leqslant \theta \leqslant 1$ and $z = 0, 1$, which are binomial. Let $g$ be a function mapping $\Theta$ to an interval $\Omega$ in a $v$-dimensional Euclidean space $(1 \leqslant v \leqslant r)$. Then, if $\hat{\theta}$ is a MLE of $\theta$, $g(\hat{\theta})$ is a MLE of $g(\theta)$. Now, $(b + c)/n$ is a maximum likelihood estimate of $2p$ (Wilks, 1962, p. 392, exercise 12.23). Applying Zehna's theorem, we see that $(b + c)/2n$ is a MLE of $p$. In this particular case, $\Theta = \{\theta\colon 0 \leqslant \theta \leqslant 1\}$, $\Omega = \{\omega\colon 0 \leqslant \omega \leqslant 1/2\}$, and $g$ simply maps $2p$ into $p$.

Next, we prove that $\hat{p}$ is an efficient estimator of $p$, that is, no other unbiased estimator has a smaller variance. We here are defining efficiency in a conventional manner, but we are aware that Rao (1973, section 5C.2) has identified some other concepts relative to efficiency. Again, we assume that we have a family of distribution functions depending on a parameter $\theta$ and that $T = \{P_\theta;\ \theta \in \Theta\}$. Let $f(z;\ \theta)$ denote the probability density function of $P_\theta$ where $Z$ is a random variable. For example, for the binomial, $Z$ would take the values 0 and 1. We now need to describe conditions under which a certain inequality holds. The inequality is

$$\sigma_\theta^2 \geqslant \frac{1}{-nE\left[\dfrac{\partial^2 \ln f(z;\theta)}{\partial \theta^2}\right]}, \tag{5}$$

which describes a lower bound to the variance of an estimator $\hat{\theta}$. ($n$ is the number of observations and ln the natural logarithm.) For this inequality to hold, a family $T$ must satisfy the four Cramer-Rao regularity conditions. (See Zacks, 1971, p. 182.) The four conditions are:

(a) $\Theta$ is either the real line, or an interval on the real line;
(b) $(\partial/\partial\theta)f(z;\ \theta)$ exists and is finite almost surely (See Loeve, 1963, p. 149) for every $\theta \in \Theta$;
(c) $\int|(\partial^i/\partial\theta^i)f(z;\ \theta)|dz < \infty$ for every $\theta \in \Theta$, and $i = 1, 2$;
(d) $E(\partial/\partial\theta \log f(z;\ \theta))^2 < \infty$ for every $\theta \in \Theta$.

Suppose $T$ is Cramer-Rao regular. Let $g$ be a real valued function on $\Theta$ and $\hat{\theta}$ be an unbiased estimator of $\theta$ having a finite variance and satisfying this fifth condition:

(e) $\int|\hat{\theta}(\partial/\partial\theta)f(z;\ \theta)|dz < \infty$.

Then inequality (5) holds. We observe that for conditions (c) and (e) the integral is technically a Lebesque integral and handles discrete distributions. In the binomial case that we are concerned with, $f(z;\ \theta)$ is given by $(2p)^z(1 - 2p)^{1-z}$, with $z$ equal to zero or one.

We wish to show that the variance of $\hat{p}$, which is equal to

$$\frac{n(2p)(1-2p)}{4n^2} = \frac{p(1-2p)}{2n},$$

has the minimum value specified by the *equality* in (5). We will then indicate that the regularity conditions are satisfied. For the right hand side of (5) we have:

$$\ln(f(z)) = z\ln(2p) + (1-z)\ln(1-2p)$$

$$\frac{\partial \ln(f(z))}{\partial p} = zp^{-1} - \frac{2(1-z)}{1-2p}$$

and

$$\frac{\partial^2 \ln(f(z))}{\partial p^2} = -zp^{-2} - 4(1-z)(1-2p)^{-2}$$

Then the expected value of this last function is:

$$-2p^{-1} - 4(1-2p)^{-2} + 8p(1-2p)^{-2},$$

since the expected value of $z$ is $2p$. This expression may be simplified by these steps:

$$\left\{\begin{array}{r} \dfrac{-2(1-2p)^2-4p+8p^2}{p(1-2p)^2} = \\ \dfrac{-2(1+4p^2-4p)-4p+8p^2}{p(1-2p)^2} = \\ \dfrac{-2(1-2p)}{p(1-2p)^2} = \\ \dfrac{-2}{p(1-2p)}. \end{array}\right. \tag{6}$$

We now multiply (6) by $-n$ and take its reciprocal, which gives $p(1-2p)/2n$ as the value of the right side of (5). This is the variance of $p$, and so $\sigma_p{}^2$ is a minimum.

It remains to show that the regularity conditions hold. We make the following observations:

(a) The first of the four regularity conditions holds, since by assumption $0 < p \leqslant 1/4$.

(b) $\partial f/\partial p = 2z(2p)^{z-1} - 2(1-z)(2p)^z(1-2p)^{-z}$ exists and is finite for every $p$, $0 < p \leqslant 1/4$.

(c) It can be seen that $\partial^2 f/\partial p^2$ also exists and is finite for every $p$, $0 < p \leqslant 1/4$. Furthermore, $f$ has only two possible values when $n = 1$. It then follows that the third regularity condition holds since $\int |(\partial/\partial p)f(z; p)|dz = \sum_{z=0}^{1} \partial f/\partial p < \infty$. In a similar fashion, condition (e) holds.

(d) Finally, the fourth condition holds since, as we saw above,

$$\left[\frac{\partial \ln(f(z))}{\partial p}\right]^2$$

exists and is finite for every $p$, $0 < p \leqslant 1/4$.

We also observe that the regularity conditions are satisfied in the more general case in which $0 < p < 1/2$. Consequently, the condition that $p \leqslant 1/4$ need not hold in order for $\hat{p}$ to be an efficient estimator of $p$. However, a large value of $p$ would be difficult to interpret in a stability study, since then the frequency of 1,0 or 0,1 responses would be greater than the frequency of consistent responses.

*Point Estimate of $x$.* Recall that $x$ is the probability of an inappropriate response, and that it is defined in terms of parameters as $x = 1/2(1 \pm (1-4p_{10})^{1/2})$. We now wish to estimate $x$. A maximum likelihood estimate of $x$ will be given by $\hat{x} = 1/2(1-(1-2(b+c)/n)^{1/2})$ providing that $0 \leqslant P_{11} \leqslant 1$, $0 \leqslant 2p \leqslant 1$, $0 \leqslant P_{00} \leqslant 1$, and $p = P_{10} = P_{01}$.*

---

*Clarification of the limits on $2p$ should be made. Earlier we made the assumption that $2p > 0$ so that the Cramer-Rao regularity conditions would be satisfied. This assumption

Define a function: $g(P_{11}, 2p, P_{00}) = \frac{1}{2}(1 - (1 - 2p)^{1/2})$, and note that $g$ is a mapping from a three-dimensional space into a two-dimensional Euclidean space or complex plane. Now $a/n$, $(b + c)/n$, and $d/n$ are MLE's of $P_{11}$, $2p$, and $P_{00}$ respectively. Substituting these values in the function defined above, we have by Zehna's theorem that $\hat{x}$ is a MLE of $x$, without any other restriction on $x$, i.e., not yet requiring $0 \leqslant x \leqslant 1$.

It is true that $\hat{x}$ can take on complex values, i.e., it can have an imaginary component. This is not a problem for the mathematics involved, but it is a problem for the application of the model to data. Suppose for a moment that $\hat{x}$ is complex. If the decision is made to accept this value as a reasonable value of $x$, then one has in effect made a decision that the L-K model does not hold, since a complex value of $x$ can hardly be interpreted as a probability. The problem that arises is that a complex value for $\hat{x}$ is possible even when the L-K model really does hold and $x$ is real. For example, suppose that the L-K model holds. Then the event $W = n$ occurs with probability greater than zero, and when it does, $\hat{x} = \frac{1}{2}(1 - \sqrt{-1}\,)$ and is complex.

Suppose we assume that $p \leqslant \frac{1}{4}$ despite any evidence to the contrary. Based on this assumption, $W/n$ is not always a MLE of $2p$ as will be shown below, and it is not necessarily true that $\hat{x}$ is a MLE of $x$. The problem to be solved is this: If we assume that $2p \leqslant \frac{1}{2}$, what is a MLE of $2p$? Once we have a MLE of $2p$, we can apply Zehna's theorem to find a MLE of $x$, to be called $\hat{x}_1$, based on the assumption that $2p \leqslant \frac{1}{2}$. In order to solve this problem, a more formal approach to maximum likelihood estimation is necessary.

Following Rao (1973, p. 353) and Zacks (1971, p. 222), let $z$ denote the realized value of a set of observations and $f(z;\ \theta)$ denote the joint density where $\theta = (\theta_1, \ldots, \theta_r)$. The vector $\theta$ is assumed to be an element of $\Theta$ where $\Theta$ is a subset of a Euclidean $r$-space. The likelihood of $\theta$ given the observations is defined as $L(\theta|z) = t f(x;\ \theta)$, where the proportionality factor $t$, $0 < t < \infty$ could depend on $z$ but is independent of $\theta$. An estimate $\hat{\theta}$ of $\theta$ is said to be maximum likelihood if

$$L(\hat{\theta}|z) = \sup_{\theta \in \Theta} L(\theta|z),$$

where sup means supremum or least upper bound.

Recall that the random variable $W$ is binomial with parameters $n$ and $2p$ where, by assumption, $p = P_{10} = P_{01}$. For convenience, set $w$

---

was necessary in order to establish that $\hat{p} = (b + c)/2n$ is efficient. Here, however, we are interested in obtaining a maximum likelihood estimate of $x$. It can be shown that when $2p = 0$, $\hat{x}$ is still a maximum likelihood estimator of $x$. Therefore, we include the possibility that $2p = 0$. The other point that needs to be made is that we are considering the problem of estimating $x$ assuming only that $P_{10} = P_{01}$. In this case, $2p$ can be greater than $\frac{1}{2}$ but must be less than or equal to 1. If, instead, we assume that the L-K model holds (i.e., $2p \leqslant \frac{1}{2}$) then $\hat{x}$ is not a maximum likelihood estimator of $x$ when $\hat{x} > \frac{1}{2}$. This point will be discussed in more detail later.

$= b + c$, that is, $w$ denotes the observed number of inconsistent responses to our hypothetical test item. The likelihood of $2p$, given the $n$ observations $a$, $b$, $c$, and $d$, is

$$L(2p|z) = (2p)^{w}(1-2p)^{n-w}. \tag{7}$$

Note that $z$, the realized value of the set of observations, is a vector of 1's and 0's. Furthermore, the number of 1's in the vector $z$ is $w$. In this instance, $\theta = 2p$ and $\Theta = \{\theta\colon\ 0 \leqslant \theta \leqslant 1\}$. As pointed out earlier, $w/n$ is a MLE of $2p$, i.e.,

$$L(w/n|z) = \sup_{\theta \in \Theta} L(\theta|z).$$

Consider $\Theta^* = \{\theta\colon\ 0 \leqslant \theta \leqslant \tfrac{1}{2}\}$. If $w/n \leqslant \tfrac{1}{2}$, then the partial derivative of $L(\theta|z)$ with respect to $\theta$ where $\theta = 2p$, vanishes at a point in $\Theta^*$. It follows that if $w/n \leqslant \tfrac{1}{2}$, then

$$L(w/n|z) = \sup_{\theta \in \Theta} L(\theta|z).$$

If we substitute $\theta$ for $2p$ in expression (7) and differentiate with respect to $\theta$, we obtain

$$w\theta^{w-1}(1-\theta)^{n-w} - (n-w)\theta^{w}(1-\theta)^{n-w-1}. \tag{8}$$

It can be shown that (8) vanishes for three values of $\theta$, namely, $\theta = 0$, $\theta = 1$ and $\theta = w/n$. Since the likelihood function is a minimum at $\theta = w/n$, it follows that if $w/n \geqslant \tfrac{1}{2}$, then $L(\theta|z)$ is (monotonically) increasing for $\theta \in \Theta^*$. Otherwise (8) would vanish in $\Theta^*$ for some point other than $\theta = 0$. Hence,

$$\sup_{\theta \in \Theta^*} L(\theta|z) = L(\tfrac{1}{4}|z).$$

We have proved the following: If one assumes that $2p \leqslant \tfrac{1}{2}$, then

$$\hat{p}_1 = \begin{cases} w/n, & \text{if } w/n < \tfrac{1}{2} \\ \tfrac{1}{2}, & \text{if } w/n \geqslant \tfrac{1}{2} \end{cases}$$

is a MLE of $2p$. To find a MLE of $p$, we consider the function $g(2p) = p$. Zehna's theorem says that

$$\hat{p}_1/2 = \begin{cases} w/2n, & \text{if } w/n < \tfrac{1}{2} \\ \tfrac{1}{4}, & \text{if } w/n \geqslant \tfrac{1}{2} \end{cases}$$

is a MLE of $p$. Finally we consider $g$ defined as

$$g(2p) = \tfrac{1}{2}(1 - (1-4p)^{1/2}).$$

Applying Zehna's theorem once more we see that

$$\hat{x}_1 = \begin{cases} \frac{1}{2}(1 - (1 - 2w/n)^{1/2}), & \text{if } w/n < \frac{1}{2} \\ \frac{1}{2}, & \text{if } w/n \geqslant \frac{1}{2} \end{cases}$$

is a MLE of $x$ assuming that $2p \leqslant \frac{1}{2}$. Most likely, both $\hat{x}_1$ and $\hat{x}$ are biased statistics for estimating $x$. The reason is that, in general, if $E(\hat{\theta}) = \theta$ then $E[g(\hat{\theta})] \neq g(\theta)$ where $E$ designates expectation. Furthermore, maximum likelihood estimates can be biased. For example, the usual statistic for estimating the correlation between two random variables with a bivariate normal distribution is a maximum likelihood, biased statistic.

Finally, we show that $\hat{x}$ is consistent. This means that $\hat{x}$ approaches $x$ in probability as $n$, the sample size, gets large. This can be stated more formally as follows. Let $\varepsilon$ be any positive real number. If

$$\lim_{n \to \infty} Pr(|\hat{x} - x| \geqslant \varepsilon) = 0,$$

then $\hat{x}$ is a consistent estimator of $x$. In other words, the probability that the absolute value of the difference of $x$ and $\hat{x}$ is greater than $\varepsilon$ approaches 0 as $n$ gets large. Furthermore, this is true for any $\varepsilon$ chosen arbitrarily close to 0. Under very general conditions, maximum likelihood statistics are consistent (see Kendall and Stuart, 1967, section 18.10). Hence, consistency is usually a less interesting property given a maximum likelihood statistic. However, several examples of inconsistent maximum likelihood estimators are known (Bahadur, 1958; Basu, 1955; Hannan, 1960; Kiefer and Wolfowitz, 1956; Neyman and Scott, 1948). For this reason, we establish that $\hat{x}$ is consistent.

It was noted earlier that $W$ is a binomial random variable with parameters $n$ and $2p$. Thus, by the law of large numbers $W/n$ approaches $2p$. Hence, $1 - 4(W/2n) \to 1 - 4p$ as $n \to \infty$ which implies

$$\sqrt{1 - 4(W/2n)} \to \sqrt{1 - 4p} \quad \text{as} \quad n \to \infty,$$

(see, Wilks, 1962, section 4.3). Thus,

$$\tfrac{1}{2}[1 - \sqrt{1 - 4(W/2n)}\,] \to \tfrac{1}{2}[1 - \sqrt{1 - 4p}] \quad \text{as} \quad n \to \infty,$$

and $\hat{x}$ is a consistent MLE of $x$.

*The Exact Probability Density Function of $\hat{X}$.* In this section we derive the probability density function of $\hat{X} = \frac{1}{2}(1 - (1 - 2W/n)^{1/2})$ which may be used, for example, in making interval estimates of $x$. That is, we want to derive a method for determining the probability that the random variable $\hat{X}$ has a particular value, say $\hat{x}$. Here we are making a distinction between random variables, e.g., $\hat{X}$, and their admissible values, e.g., $\hat{x}$. Symbolically, we want to determine $Pr[\hat{X} = \hat{x}]$. We will demonstrate how $Pr[\hat{X} = \hat{x}]$ can be expressed in terms of $2p$ where, as before, $p = P_{01} = P_{10}$. As noted earlier

$$\hat{X} = \tfrac{1}{2}(1 - (1 - 2W/n)^{1/2}). \tag{9}$$

We first solve for $W$ as follows $2\hat{X} - 1 = -1(1 - 2W/n)^{1/2}$, and $(2\hat{X} - 1)^2 = 1 - 2W/n$, or $2W/n = 1 - 4\hat{X}^2 + 4\hat{X} - 1$. Then $W = 2n(\hat{X} - \hat{X}^2)$. This means that the event $W = w$ corresponds to the event $\hat{X} = \hat{x}$ for one and only one value $\hat{x}$ of $\hat{X}$. Equivalently, $Pr[W = w] = Pr[\hat{X} = \hat{x}]$. Verbally, the probability of observing $w$ of $n$ randomly chosen examinees who answer the test item correctly on one occasion and incorrectly on the other is equal to the probability of calculating $\hat{X}$ to be $\hat{x}$ since the value $\hat{x}$ corresponds to exactly one possible value $w$ of the random variable $W$. To be more explicit, recall that $W$ can have one of $n + 1$ possible values, namely, 0, 1, 2, . . . $n$. If, for example, $W = 0$ (all $n$ of the examinees give a consistent response to the test item) then from (9) $\hat{X} = \tfrac{1}{2}$. Also, as was shown above, $\hat{X} = \tfrac{1}{2}$ implies that $W = 0$. Consequently, $Pr[W = 0] = (1 - 2p)^n = Pr[\hat{X} = \tfrac{1}{2}]$. For any value $w$ of $W$:

$$Pr[W = w] = \binom{n}{w}(2p)^w(1 - 2p)^{n-w} = Pr[\hat{X} = \hat{x}], \tag{10}$$

where $\hat{x}$ is the value of $\hat{X}$ corresponding to $w$. Since $w = 2n(\hat{x} - \hat{x}^2)$ we may write $Pr[\hat{X} = \hat{x}]$ as

$$\binom{n}{2n(\hat{x} - \hat{x}^2)}(2p)^{2n(\hat{x}-\hat{x}^2)}(1 - 2p)^{n-2n(\hat{x}-\hat{x}^2)} \tag{11}$$

by substituting $2n(\hat{x} - \hat{x}^2)$ for $w$ in (10). Expression (11) is the probability density function of $\hat{X}$ and gives in terms of $2p$ the probability that $\hat{X}$ will have a particular admissible value $\hat{x}$ which is the desired result.

*Point Estimate of* $k$. Recall that $k$, the proportion of examinees who know the test item, is given by

$$k = \tfrac{1}{2} + \frac{2P_{1.} - 1}{2(1 - 4p)^{1/2}} = \tfrac{1}{2} + \frac{2(P_{11} + p) - 1}{2(1 - 4p)^{1/2}}$$

In simple terms, the theorem by Zehna described above says that if we substitute a maximum likelihood estimate of $P_{1.}$ and $p$ in the expression for $k$, then we obtain a maximum likelihood estimate of $k$. Since $(a + b)/n$ and $(b + c)/2n$ are MLE's of $p_{1.}$ and $p$, as noted earlier, we have that

$$\hat{k} = \tfrac{1}{2} + \frac{2((2a + b + c)/2n) - 1}{2(1 - 4((b + c)/2n))^{1/2}} = \tfrac{1}{2} + \frac{(2a + b + c)/n - 1}{2(1 - 2((b + c)/n))^{1/2}}$$

is a MLE of $k$.

As was the case with $\hat{x}$, the event that $\hat{k}$ has a complex value occurs with probability greater than zero even when the L-K model holds, in which case $k$ is no longer a MLE of $k$ when we assume the model holds. If a decision is made that the value of $k$ is indeed a complex number (has an imaginary component) then the decision has been made that the L-K model does not hold since $k$ can hardly be interpreted as a proportion. Consequently, there may be situations in which one is unwilling to accept complex numbers as estimates of $k$. One approach to this problem is to consider the estimation of $k$ under the assumption that $p = P_{10} = P_{01}$ and that $2p \leq \frac{1}{2}$, which precludes complex numbers as appropriate estimates of $k$. In this case, $\hat{k}$, as defined above, is no longer a MLE of $k$. To find a MLE of $k$, to be called $\hat{k}_1$, it is sufficient, according to the theorem by Zehna, to find a MLE of $2p$ and $P_{11}$ assuming that $p = P_{10} = P_{01}$ and $2p \leq \frac{1}{2}$. It will be shown that

$$\hat{P}_{11} = \begin{cases} a/n & \text{if } (b+c)/n < \frac{1}{2} \\ a/2(n-b-c) & \text{if } (b+c)/n \geq \frac{1}{2} \end{cases}$$
$$2\hat{p} = \begin{cases} (b+c)/n & \text{if } (b+c)/n < \frac{1}{2} \\ \frac{1}{2} & \text{if } (b+c)/n \geq \frac{1}{2} \end{cases} \qquad (12)$$

yield a maximum likelihood estimator for the problem of *simultaneously* estimating $P_{11}$ and $2p$ assuming that $p = P_{10} = P_{01}$ and $2p < \frac{1}{2}$. Before continuing, however, we will give a non-technical explanation of (12) as an estimate of $P_{11}$.

The estimate of $P_{11}$ with $a/2(n - b - c)$ when $(b + c)/n \geq \frac{1}{2}$ is counterintuitive since it would seem that the estimate of $P_{11}$ should not depend in any particular way on the values $b$ and $c$ nor on the admissible values of $p$. As a brief explanation, we first point out that if one adopts a maximum likelihood estimation approach to estimating $P_{11}$ and $2p$, then we are in effect searching for a point in the admissible region of a plane (the plane being the domain of an appropriate likelihood function) such that for this point the likelihood function is a maximum. If, by assumption, $2p < \frac{1}{2}$ and $(b + c)/n \geq \frac{1}{2}$ then the problem of maximizing the likelihood function reduces to finding an appropriate point in a line in the plane. Moreover, this line determines the form of the likelihood function so that the point at which the likelihood equation is a maximum can be determined. It is the form of the likelihood equation for the points on this line that accounts for the dependence of the estimator $\hat{P}_{11}$ on the values of $b$ and $c$.

If we use (12) as maximum likelihood estimators of $P_{11}$ and $2p$, then according to the theorem by Zehna, we obtain a MLE of $k$ by substituting the corresponding estimates of $P_{11}$ and $2p$ in the expression

$$k = \frac{1}{2} + \frac{2P_{11} + 2p - 1}{2(1 - 4p)^{1/2}} \qquad (13)$$

Suppose, for example, that $(b + c)/n \geq \frac{1}{2}$. To estimate $k$, we would substitute $a/2(n - b - c)$ for $P_{11}$ and $\frac{1}{2}$ for $2p$ in expression (13). However, such a procedure would result in division by zero since $1 - 4p = 1 - 2(2p) = 1 - 2(\frac{1}{2}) = 0$. If, of course, $(b + c)/n < \frac{1}{2}$ and we substitute $(b + c)/n$ for $2p$ in (13), then the problem of division by zero does not arise. To avoid division by zero we might choose some positive number close to zero, say $\varepsilon$, and assume that $2p \leqslant \frac{1}{2} - \varepsilon$. If, for example, we choose $\varepsilon$ to be .01, then the assumption that $2p < .5$ is nearly equivalent to assuming that $2p < .5 - .01 = .49$. For a positive $\varepsilon$ less than $\frac{1}{2}$, it is shown below that

$$\hat{P}_{11} = \begin{cases} a/n & \text{if } (b + c)/n < \frac{1}{2} - \varepsilon \\ a(\frac{1}{2} + \varepsilon)/(n - b - c) & \text{if } (b + c)/n \geqslant \frac{1}{2} - \varepsilon \end{cases}$$
$$2\hat{p} = \begin{cases} (b + c)/n & \text{if } (b + c)/n < \frac{1}{2} - \varepsilon \\ \frac{1}{2} - \varepsilon & \text{if } (b + c)/n \geqslant \frac{1}{2} - \varepsilon \end{cases} \tag{14}$$

provides a MLE for the problem of simultaneously estimating $P_{11}$ and $2p$ under the assumption that $2p < \frac{1}{2} - \varepsilon$. If, for example, $(b + c)/n < \frac{1}{2} - \varepsilon$ then an application of Zehna's theorem implies that we estimate $k$ to be

$$\frac{1}{2} + \frac{2a/n + (b + c)/n - 1}{2(1 - 2((b + c)/n))^{1/2}}.$$

If $(b + c)/n \geqslant \frac{1}{2} - \varepsilon$ then the estimate of $k$ would be

$$\frac{1}{2} + \frac{2a(\frac{1}{2} + \varepsilon)/(n - b - c) + \frac{1}{2} - \varepsilon - 1}{2(1 - 2(\frac{1}{2} - \varepsilon))^{1/2}}$$
$$= \frac{1}{2} + \frac{2a(\frac{1}{2} + \varepsilon)/(n - b - c) - \frac{1}{2} - \varepsilon}{2(2\varepsilon)^{1/2}}$$

We now give a proper verification that (12) is a MLE of $P_{11}$ and $2p$ assuming that $p = P_{10} = P_{01}$ and $2p \leqslant \frac{1}{2}$ after which we consider the problem of estimating $P_{11}$ and $2p$ when by assumption $2p \leqslant \frac{1}{2} - \varepsilon$. Since we have already assumed that $P_{10} = P_{01}$, this particular situation can be described with a trinomial distribution. That is, each student falls into one of three categories: (1) he is correct on both occasions, (2) he is inconsistent in his response, or (3) he is incorrect on both occasions. The corresponding probabilities are $P_{11}$, $2p = P_{10} + P_{01}$ and $P_{00} = 1 - P_{11} - 2p$. Note that the $\hat{p}_1$ given above does not necessarily apply to the present situation, since the derivation of $\hat{p}_1$ was based on a binomial distribution.

To be consistent with the notation used for describing MLE's, set $\theta_1 = P_{11}$ and $\theta_2 = 2p$. The likelihood function is

$$L(\theta|z) = \theta_1^{a}\theta_2^{w}(1 - \theta_1 - \theta_2)^{n-a-w} \tag{15}$$

where $\theta = (\theta_1, \theta_2)$, $w = b + c$, $z$ denotes the realized values of the random variables, and $a$, $b$, and $c$ are observed frequencies. When

$0 \leqslant \theta_1 \leqslant 1$ and $0 \leqslant \theta_2 \leqslant 1$ then $a/n$ and $w/n$ are MLE's of $\theta_1$ and $\theta_2$, respectively, as noted earlier. That is,

$$L((a/n,\ w/n)|z) = \sup_{\theta \in \Theta} L(\theta|z)$$

where $\Theta = \{(\theta_1, \theta_2)\colon\ 0 \leqslant \theta_1 \leqslant 1,\ 0 \leqslant \theta_2 \leqslant 1,\ \theta_1 + \theta_2 \leqslant 1\}$ Let $\Theta^* = \{(\theta_1, \theta_2)\colon\ 0 \leqslant \theta_1 \leqslant 1,\ 0 \leqslant \theta_2 \leqslant ½,\ \theta_1 + \theta_2 \leqslant 1\}$. The elements of the set $\Theta$ correspond to the points in the triangle EIH of Figure 1.

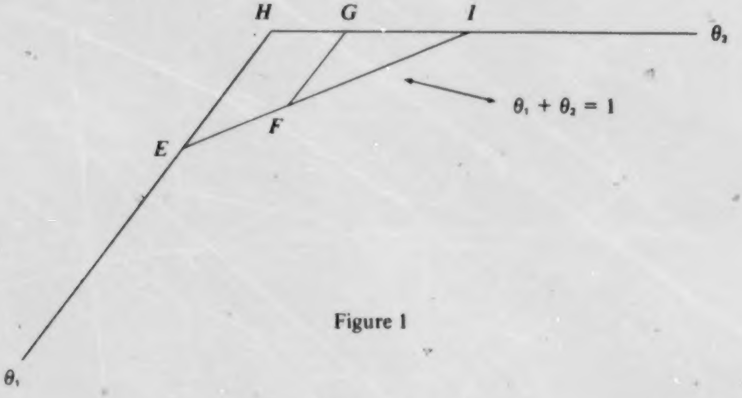

Figure 1

$\Theta$ is the admissible part of the plane for the general trinomial distribution. The elements of $\Theta^*$ are the points in the quadrilateral *EFGH* where the point $H$ is $\theta = (0,\ 0)$, $G$ is $\theta = (0,\ ½)$, $I$ is $\theta = (0,\ 1)$, $E$ is $\theta = (1,\ 0)$ and $F$ is $\theta = (½,\ ½)$. We wish to show that

$$\hat{\theta} = \begin{cases} (a/n,\ w/n), & \text{if } w/n < ½ \\ (a/2(n-w),\ ½), & \text{if } ½ \leqslant w/n \leqslant 1 \end{cases}$$

is a MLE of $\theta$ for $\theta \in \Theta^*$.

We first consider whether or not a MLE of $\theta$ exists—a fact that cannot always be taken for granted (see Rao, 1973, section 5d.3). Since the $\sup_{\theta \in \Theta}$ exists for (15), expression (15) is bounded, and therefore is bounded for $\theta \in \Theta^*$, which implies that $\sup_{\theta \in \Theta^*} L(\theta|z)$ exists (Rudin, 1964, theorem 1.36). Thus, a MLE exists.

If $w/n < ½$ then $\hat{\theta} = (a/n,\ w/n)$ is a MLE of $\theta$ since, in particular, $\partial L/\partial\theta_1$ and $\partial L/\partial\theta_2$ vanish (see Wilks, 1962, p. 392). However, if $w/n \geqslant ½$, then the partial derivatives do not vanish for any interior point of the quadrilateral *EFGH* in Figure 1. To see this, we differentiate (15) with respect to $\theta_1$ to obtain

$$\theta_2^w[a\theta_1^{a-1}(1 - \theta_1 - \theta_2)^{n-a-w} - \theta_1^a(n - a - w)(1 - \theta_1 - \theta_2)^{n-a-w-1}].$$

Differentiating (15) with respect to $\theta_2$ we have that

$$\theta_1^a[w\theta_2^{w-1}(1 - \theta_1 - \theta_2)^{n-a-w} - \theta_2^w(n - a - w)(1 - \theta_1 - \theta_2)^{n-a-w-1}].$$

If $0 < \theta_1 < 1$ and $0 < \theta_2 < 1$, then the two expressions above can be equal to zero if and only if $\theta_1 = a/n$ and $\theta_2 = w/n$. Thus, there exists a $\theta$, say $\theta_0$, on the boundary of the quadrilateral *EFGH* such that $L(\theta_0|z) \geqslant L(\theta|z)$ for any $\theta \in \Theta^*$ by definition of sup. If

$$\theta \in \{(\theta_1, \theta_2): \theta_1 + \theta_2 = 1\}$$

or if $\theta$ is of the form $(0, \theta_2)$ or $(\theta_1, 0)$ then the right hand side of (15) is zero. But there is always a point $\theta$ for which (15) is not zero. Thus, we arrive at the conclusion that the maximum of (15) must occur at a point $\theta$ where $\theta$ is of the form $\theta = (\theta_1, \frac{1}{2})$ and $0 \leqslant \theta_1 \leqslant \frac{1}{2}$. Substituting $\frac{1}{2}$ for $\theta_2$ in expression (15) gives

$$\theta_1^a(\tfrac{1}{2})^w(1 - \theta_1 - \tfrac{1}{2})^{n-a-w} = \theta_1^a(\tfrac{1}{2})^w(\tfrac{1}{2} - \theta_1)^{n-a-w}. \qquad (16)$$

We are now left with the problem of choosing a value for $\theta_1$ which maximizes (16). Differentiating (16) with respect to $\theta_1$ and setting the result equal to zero gives

$$(\tfrac{1}{2})^w[(a\theta_1^{a-1})(\tfrac{1}{2} - \theta_1)^{n-a-w} - (n - a - w)\theta_1^a(\tfrac{1}{2} - \theta_1)^{n-a-w-1})] = 0$$

Thus,

$$a\theta_1^{a-1}(\tfrac{1}{2} - \theta_1)^{n-a-w} = (n - a - w)\theta_1^a(\tfrac{1}{2} - \theta_1)^{n-a-w-1}. \qquad (17)$$

Dividing both sides of (17) by $\theta_1^{a-1}(1 - \theta_1)^{n-a-w-1}$ gives $a(\frac{1}{2} - \theta_1) = (n - a - w)\theta_1$. Hence, $(n - w)\theta_1 = a/2$ and thus, $\theta_1 = a/2(n - w)$. This completes the proof. Note that if $w = n$ then (15) is a maximum for any $\theta \in \{(\theta_1, \theta_2): 0 \leqslant \theta_1 \leqslant \frac{1}{2}, \theta_2 \leqslant \frac{1}{2}\}$.

If we assume that $0 \leqslant \theta_1 \leqslant \frac{1}{2} - \varepsilon$ rather than $0 \leqslant \theta_1 \leqslant \frac{1}{2}$ then it can be shown that the likelihood function is a maximum at a point of the form $\theta = (\theta_1, \frac{1}{2} - \varepsilon)$. It follows that (14) is a maximum likelihood estimator of $P_{11}$ and $p$. The details of the argument are omitted since they are exactly the same as the case $0 \leqslant \theta_1 \leqslant \frac{1}{2}$.

To find the distribution of $\hat{K}$, the random variable corresponding to the statistic $\hat{k}$, we first derive the joint distribution of $\hat{X}$ and $\hat{K}$. The technique used is essentially the same as the one for finding the distribution of $\hat{X}$ (see Hogg and Craig, 1970, section 4.2). Recall that $A$ is the random variable that denotes the number of students who answer the test item correctly on both occasions. $A$ has a binomial distribution with parameters $n$ and $P_{11}$. Now the transformation

$$\hat{X} = \tfrac{1}{2}[1 - \sqrt{1 - 4(W/2n)}\,] \tag{18}$$

and

$$\hat{K} = \tfrac{1}{2} + \frac{2(A/n + W/2n) - 1}{2\sqrt{1 - 4(W/2n)}} \tag{19}$$

is one-to-one. To show this, we first note that $W = 2n(\hat{X} - \hat{X}^2)$ as shown before. Next we solve for $A$. From (19), $(\hat{K} - \tfrac{1}{2})(2(1 - 2W/n)^{1/2}) = 2(A/n + W/2n) - 1$.

Hence, $\tfrac{1}{2}((\hat{K} - \tfrac{1}{2})2(1 - 2W/n)^{1/2} + 1) = A/n + W/2n$.

Thus, $\tfrac{1}{2}((\hat{K} - \tfrac{1}{2})2(1 - 2W/n)^{1/2} + 1) - W/2n = A/n$.

Whence, $A = (n/2)((\hat{K} - \tfrac{1}{2})2(1 - 2W/n)^{1/2} + 1) - \tfrac{1}{2}W$.

Substituting $2n(\hat{X} - \hat{X}^2)$ for $W$ gives

$$A = (n/2)((\hat{K} - \tfrac{1}{2})2(1 - 4(\hat{X} - \hat{X}^2))^{1/2} + 1) - n(\hat{X} - \hat{X}^2)$$

But the joint p.d.f. of $A$ and $W$ is

$$\frac{n!P_{11}^{a}(2p)^{w}(1 - 2p - P_{11})^{n-a-w}}{a!w!(n - a - w)!}. \tag{20}$$

As usual, lower case letters denote a value of a random variable. By substituting $2n(\hat{x} - \hat{x}^2)$ for $w$ and $n[(\hat{k} - \tfrac{1}{2})(1 - 4(\hat{x} - \hat{x}^2))^{1/2} + \hat{x}^2 - \hat{x} + \tfrac{1}{2}]$ for $a$ in (20), we have the joint p.d.f. of $\hat{X}$ and $\hat{K}$. What this means is that the event $A = a$ and $W = w$ corresponds to one and only one pair of values of $\hat{K}$ and $\hat{X}$, say $\hat{x}$ and $\hat{k}$. Moreover, $Pr(A = a, W = w) = Pr(\hat{X} = \hat{x}, \hat{K} = \hat{k})$.

From (19) and (20) we see that the possible values of $\hat{K}$ are determined by the value of $\hat{X}$. The reason is that a given value of $W$ determines the possible values of $A$. Thus, $\hat{X}$ and $\hat{K}$ are dependent.

Let $f_2$ denote the joint p.d.f. of $\hat{X}$ and $\hat{K}$. Then the p.d.f. of $\hat{K}$ is $\sum_x f_2(\hat{x}, \hat{k})$ where the sum is over all possible values of $\hat{X}$ for a given value of $\hat{K}$. From (18) and (19) it appears that there is one and only one value of $\hat{X}$ for a given value of $\hat{K}$ since the values of $A$ and $W$ are integers. However, it is not completely clear that this is always the case. Therefore, a certain amount of caution must be exercised if one is interested in using the exact *p.d.f.* of $\hat{K}$.

*The Asymptotic Distribution of $\hat{X}$ and $\hat{K}$.* Unfortunately, the exact p.d.f. of $\hat{X}$ alone or $\hat{K}$ alone in the L-K model is not very convenient to use. For this reason, we examine the asymptotic distribution of both $\hat{X}$ and $\hat{K}$.

Wilks (1962, section 9.3) states the following theorem: Suppose $V_{1m}, \ldots, V_{jm}$, $m = 1, \ldots, n$ is a sample from a $j$-dimensional distribution with finite means, say $\mu_j$, and positive definite variance-covariance matrix $\|\sigma_{im}\|(i, m = 1, \ldots, j)$. Let $g(V_1, \ldots, V_j)$ be a function and suppose the first derivatives of $g$, say $\partial g/\partial V_i = g_i$, $i = 1, \ldots, j$ exist at all points in some neighborhood of $(\mu_1, \ldots, \mu_j)$ and let $g_i^\circ = g_i(\mu_1, \ldots, \mu_j)$. Then if at least one of the $g_i^\circ$ is $\neq 0$, $g(\bar{V}_1, \ldots, \bar{V}_j)$ is asymptotically normal with mean $g(\mu_1, \ldots, \mu_j)$ and variance $1/n \sum_{l,m=1}^{j} \sigma_{lm} g_l^\circ g_m^\circ$ where $\bar{V}_i = 1/n \sum_{m=1}^{n} V_{im}$.

In terms of $\hat{X}$, $j = 1$ and $\bar{V}_1 = A/n$, where, as before, $A$ is the random variable corresponding to a correct response to the test item on both occasions and where the observed value of $A$ is $a$. To see why $\bar{V}_1 = A/n$ is appropriate, note that $A/n$ is the sample mean based on a random sample of $n$ observations from a population with a binomial *p.d.f.* with parameters 1 and $P_{11}$, i.e., $A/n$ estimates $P_{11}$. As for $\hat{K}$, $j = 2$, $\bar{V}_1 = A/n$ and $\bar{V}_2 = W/n$. For $\hat{X}$ we assume that $p \neq 1/4$. If $p = 1/4$, $g_1$ does not exist at all points in any neighborhood of $2p$ since, in particular, $g_1$ is not defined at the point $p = 1/4$. If we assume that $p > 0$, then $g_1 \neq 0$. As for $\hat{K}$, it is sufficient to assume that $p \neq 1/4$ and $p > 0$. We also must verify that the variance-covariance matrix of $A$ and $W$ is positive definite. A necessary and sufficient condition that the variance-covariance matrix be positive definite is that there exists no linear dependence between $A$ and $W$ (Wilks, 1962, section 3.5). That is, it is impossible to find $c_1$ and $c_2$, not both zero, such that $Pr \cdot (c_1 A + c_2 W = c_3) = 1$ for some constant $c_3$. Since we are in effect working with a trinomial distribution, it can be seen that there is no linear dependence between $A$ and $W$ except when $P_{00} = 1$.

This line of argument leads to the conclusion that both $\hat{X}$ and $\hat{K}$ are normally distributed asymptotically, and thus, indicates that with reasonably large sample sizes we may interpret either $\hat{X}$ or $\hat{K}$ as a normally distributed variable. Moreover, the asymptotic expected value of $\hat{X}$ is $x$ and the expected value of $\hat{K}$ is $k$.

*Modifications of the Lazarsfeld-Kendall Turnover Model.* We describe two modifications of the L-K model that apparently were first suggested and investigated by Wilcox (1976). In the L-K model, the probability of an inappropriate response on the first occasion is set equal to the probability of an inappropriate response on the second occasion. An inappropriate response, however, may be regarded as one of two kinds: knowing but answering incorrectly or not knowing and answering correctly. We now describe two models that result when we set the probability of one of these two types of inappropriate responses equal to zero. When we assume that an examinee who does not know the answer guesses the answer with probability zero, then we have

| | | Second Occasion | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| First Occasion | 1 | $k(1-x)^2$ | $kx(1-x)$ | $k(1-x)$ |
| | 0 | $kx(1-x)$ | $(1-k)+kx^2$ | $kx+(1-k)$ |
| | | $k(1-x)$ | $kx+(1-k)$ | 1 |

as the model. Here $x$ is the (non-zero) probability of answering incorrectly when the examinee knows the answer, and $k$ is the proportion of examinees who know the answer on both occasions. Note that exact symmetry holds for the parameters of this model. We can solve for $x$ and $k$ in terms of parameters, thus,

$$x = p(P_{11} + p)^{-1}$$

$$k = P_{11}^{-1}(P_{11} + p)^2$$

In these equations, $p = P_{10}$ as before. A solution for $x$ assumes $P_{00} < 1$. If $P_{00} = 1$, then we take $k = 0$.

This model is similar to a model by Marks and Noll (1967); their model was developed for the case of pre and post testing associated with instruction, whereas ours is for the case of testing on two occasions when no systematic learning is assumed to take place during the interval. The model we describe may be useful for items for which the probability of guessing the correct answer is zero or very close to zero. Completion items that require specific information which is not generally available may be an illustration.

To obtain a MLE of $\hat{x}$ and $\hat{k}$ it is sufficient, according to the theorem by Zehna, to substitute a MLE of $p$ and $P_{00}$ in the expression for $\hat{x}$ and $\hat{k}$. We have then that

$$\hat{x} = (b + c)/(2a + b + c)$$

and

$$\hat{k} = (2a + b + c)^2/4an$$

are a MLE of $x$ and $k$ respectively.

The exact probability density function of $x$ and $k$ can also be derived. (See Wilcox, 1976, Chapter 2.) As was the case for the L-K model, the form of the density functions is quite involved; consequently, we merely state that the distributions of the statistics for estimating $x$ and $k$ (the probability density functions of $\hat{x}$ and $\hat{k}$) are asymptotically normal. The details of the proof are exactly the same as they were for the L-K model and are therefore omitted.

We now consider the second modification of the L-K model which is given by assuming that the probability of knowing the answer to the item and getting it wrong is zero. We define $y$ to be the probability of getting the item right given that the examinee does not know the answer. In contrast to the first modification of the L-K model, we assume that $y$ is greater than zero. The resulting fourfold table of probabilities is as follows:

| | | Second Occasion | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| First Occasion | 1 | $k + (1 - k)y^2$ | $(1 - k)y(1 - y)$ | $(1 - k)y + k$ |
| | 0 | $(1 - k)y(1 - y)$ | $(1 - k)(1 - y)^2$ | $(1 - k)(1 - y)$ |
| | | $(1 - k)y + k$ | $(1 - k)(1 - y)$ | 1 |

This modification may be appropriate for relatively easy multiple choice items associated with an elementary instructional program that is designed to teach basic facts and terms but not complicated relationships among principles. Solving for $y$ and $k$, we have that

$$y = \frac{p}{p + P_{00}}$$

$$k = 1 - \frac{(p + P_{00})^2}{P_{00}}$$

Applying the theorem by Zehna, we have that

$$\hat{y} = \frac{(b + c)/2n}{(b + c)/2n + d/n}$$

and

$$\hat{k} = 1 - \frac{((b + c)/2n + d/n)^2}{d/n}$$

are a MLE of $y$ and $k$ respectively when $p = P_{10} = P_{01}$ and when $2p$, $P_{11}$, and $P_{00}$ can have any value between zero and one inclusive with the restriction that $2p + P_{11} + P_{00} = 1$. We merely state that the distributions of $\hat{y}$ and $\hat{k}$ are asymptotically normal.

**Table of Symbols**

$x$ = the probability that a randomly chosen examinee gives an inappropriate response to the test item. $x$ is a population parameter.

$P_{11}$ = the population parameter denoting the probability that a randomly chosen examinee answers correctly on both occasions.

$P_{10}$ = a population parameter denoting the probability that a randomly chosen examinee gives a correct-incorrect response.

$P_{01}$ = the population parameter denoting the probability that a randomly chosen examinee gives an incorrect-correct response.

$P_{00}$ = the population parameter denoting the probability that a randomly chosen examinee gives an incorrect response on both occasions.

$k$ = the population parameter denoting the proportion of examinees who know the answer to the test item.

$A$ = the random variable that corresponds to the number of examinees who answer the test item correctly on both occasions.

$a$ = the value of the random variable $A$ which is observed for a particular sample of examinees. That is, $a$ is the number of examinees who answer the test item correctly on both occasions.

$B$ = the random variable corresponding to the number of examinees giving a correct-incorrect response to the test item.

$b$ = the value of the random variable $B$. Equivalently, $b$ is the observed number of examinees giving a correct-incorrect response for a particular sample of examinees.

$C$ = the random variable corresponding to the number of examinees who give an incorrect-correct response.

$c$ = the observed value of the random variable $C$.

$D$ = the random variable corresponding to the number of examinees who give an incorrect-incorrect response.

$d$ = the observed value of $D$.

$p = P_{10} = P_{01}$, where $P_{10} = P_{01}$ is the assumption of exact symmetry in the L-K model.

$W = B + C$, the random variable corresponding to the number of examinees who give an inconsistent response.

$w$ = the observed value of $W$ where $w = b + c$.

$\hat{p} = (b + c)/2n$, the observed value used to estimate $p$. Note that $\hat{p}$ is the observed value of the random variable $W/2n$.

$\hat{x} = \frac{1}{2}(1 - (1 - 2(b + c)/n)^{1/2})$. $\hat{x}$ is the observed value of the random variable $\hat{X}$.

$\hat{X} = \frac{1}{2}(1 - (1 - 2W/n)^{1/2})$. The symbol "$\hat{}$" is used so as to make a distinction between the values of the random variable $\hat{X}$, namely $\hat{x}$, and the parameter $x$.

$\theta$ = an arbitrary parameter which may be vector valued, i.e., it may represent several parameters.

$z$ = an arbitrary realized value of a set of observations. For example, $z$ may designate a sequence of zeroes and ones.

$\Theta$ = the set of admissible values of $\theta$. Also, it is a subset of a Euclidean $r$-space, $r \geq 1$.

$f$ = a probability density function.

$L(\theta|z)$ = the likelihood function of $\theta$.

$\hat{p}_1$ = an estimate of $2p$ assuming $2p \leq \frac{1}{2}$.

$\hat{x}_1$ = observed value used to estimat assuming $2p \leq \frac{1}{2}$.

$Pr$ = probability of.

$P_1 = P_{11} + P_{10}$

$\hat{K}$ = the random variable $\frac{1}{2} + \dfrac{2(A/n + W/2n) - 1}{2(1 - 4(W/2n))^{1/2}}$

$\hat{k}$ = the observed value of the random variable $\hat{K}$. The symbol "$\hat{\ }$" is used to distinguish $\hat{k}$ from the parameter $k$.

$\hat{P}_{11}$ = estimate of $P_{11}$. When $(b + c)/n < \frac{1}{2}$, $a/n$ is $\hat{P}_{11}$. When $(b + c)/n \geqslant \frac{1}{2}$, $a/2(n - b - c)$ is $\hat{P}_{11}$.

$\varepsilon$ = a positive real number.

$\Theta^*$ = a subset of $\Theta$.

## Chapter 5

# NEW METHODS FOR STUDYING EQUIVALENCE

**RAND R. WILCOX**

In this paper we describe three probability models that may be useful in equivalence studies of achievement test items. We begin by describing a model based on a modification of the Lazarsfeld-Kendall turnover model that was suggested by Goodman and Kruskal (1959). The other two models are a modification of the Goodman-Kruskal model and will be described in a subsequent section.

We assume that there are two test items each of which is answered by the same sample of subjects randomly chosen from some population of potential examinees. Each item is scored 1 for a correct or 0 for an incorrect response. Furthermore, we assume that there are two classes of people; namely, those who know the answer to both items and those who do not. We further assume that an examinee may make an inappropriate response—either a correct response when he does not know or an incorrect response when he does. We begin by deriving the probability of an inappropriate response in terms of population parameters. We will also derive an expression for the probability that a randomly chosen person answers the two questions similarly and an expression for the proportion of people in the population of potential examinees who know the answer to both items. Point and interval estimates for the parameters will be developed.

Let $k$ denote the proportion of subjects who know the answer to both items. Let $x_1$ denote the probability of an inappropriate response to item one and $x_2$ the probability of an inappropriate response to item two. By an inappropriate response we mean that an examinee responds correctly when he does not know the answer or incorrectly when he does. We assume that $x_1 \leq 1/2$, $x_2 \leq 1/2$ and that the probability of an inappropriate response to item one is independent of the probability of an inappropriate response to item two. The fourfold table of probabilities corresponding to the possible outcomes is given in Table 1.

**Table 1**

Item 2

| | | 1 | 0 | |
|---|---|---|---|---|
| Item 1 | 1 | $P_{11} = k(1 - x_1)(1 - x_2) + (1 - k)x_1x_2$ | $P_{10} = k(1 - x_1)x_2 + (1 - k) x_1(1 - x_2)$ | $k(1 - x_1) + (1 - k)x_1$ |
| | 0 | $P_{01} = kx_1(1 - x_2) + (1 - k) \cdot (1 - x_1)x_2$ | $P_{00} = kx_1x_2 + (1 - k) \cdot (1 - x_1) \cdot (1 - x_2)$ | $kx_1 + (1 - k) \cdot (1 - x_1)$ |
| | | $k(1 - x_2) + (1 - k)x_2$ | $kx_2 + (1 - k) \cdot (1 - x_2)$ | 1 |

Here, $P_{11} + P_{10} + P_{01} + P_{00} = 1$. Note that $P_{11}$, $P_{10}$, $P_{01}$, $P_{00}$, $x_1$, $x_2$, and $k$ are parameters. What we observe is:

Item 2

| | | 1 | 0 | |
|---|---|---|---|---|
| Item 1 | 1 | $a$ | $b$ | $a + b$ |
| | 0 | $c$ | $d$ | $c + d$ |
| | | $a + c$ | $b + d$ | $n$ |

where $a + b + c + d = n$ is the number of examinees answering both items.

In order to estimate $x_1$, $x_2$ and $k$ we first need to express these parameters in terms of the $P_{ij}$'s ($i = 0, 1$; $j = 0, 1$). From Table 1, it can be shown that

$$\begin{cases} x_1 = \dfrac{P_{1.} - k}{1 - 2k} \\ x_2 = \dfrac{P_{.1} - k}{1 - 2k} \\ k = \tfrac{1}{2}(1 \pm (1 - 4r)^{1/2}) \text{ where} \\ r = \dfrac{(P_{11} - P_{1.}P_{.1})}{(1 - 2(P_{10} + P_{01}))} \end{cases} \tag{1}$$

Note that in order for $k$ to be real, $r$ must be $\leqslant \frac{1}{4}$. In terms of $r$ we have:

$$\begin{cases} x_1 = \frac{1}{2} \mp \dfrac{2P_{1.} - 1}{2(1 - 4r)^{1/2}} \quad \text{and} \\[2ex] x_2 = \frac{1}{2} \mp \dfrac{2P_{.1} - 1}{2(1 - 4r)^{1/2}} \end{cases} \tag{2}$$

Goodman and Kruskal (1959, p. 151) impose the additional restriction that $(P_{1.} - \frac{1}{2})(P_{.1} - \frac{1}{2}) \geqslant 0$ so that the choice of sign for both $x_1$ and $x_2$ is the same. It can be shown that if $P_{10} = P_{01}$, then the model reduces to the Lazarsfeld-Kendall turnover model described in the previous chapter on stability. Goodman and Kruskal summarize the necessary and sufficient conditions for the model to hold as follows:

$$\begin{cases} 0 \leqslant \dfrac{P_{11} - P_{1.}P_{.1}}{1 - 2(P_{10} + P_{01})} \leqslant \text{Min}(P_{1.}P_{0.},\ P_{.1}P_{.0}) \\[2ex] (P_{1.} - \frac{1}{2})(P_{.1} - \frac{1}{2}) \geqslant 0 \end{cases} \tag{3}$$

*Point Estimates of the Parameters.* Our goal is to determine a "reasonable" approach to the problem of estimating the parameters $x_1$, $x_2$ and $k$ based on the observations. We assume that sampling is from an infinite population or from a finite population with replacement and consequently that the multinomial is the proper distribution function. Under this assumption, it will be shown that to estimate $P_{11}$ with $a/n$, $P_{10}$ with $b/n$, $P_{01}$ with $c/n$, and $P_{00}$ with $d/n$ and to substitute these estimates in the expression for $x_1$, $x_2$ and $k$ provides a maximum likelihood estimate of the parameters. A maximum likelihood estimator (MLE) of a parameter is one such that of all possible values that the parameter could possibly take, it has the value which makes the observations most likely. We begin by describing a theorem that Zehna (1966) developed.

In simple terms, Zehna has shown that, under very general conditions, if $\hat{\theta}$ is a MLE of a parameter $\theta$ where $\theta$ is vector valued, then for any function $g$, $g(\hat{\theta})$ is a MLE of $g(\theta)$ provided $g$ is a mapping from Euclidean $u$-space into a Euclidean $v$-space where $u \geqslant v$. It has been known for some time that $g(\hat{\theta})$ is a MLE of $g(\theta)$ provided $g$ is one-to-one, that is, $g(\theta_1) = g(\theta_2)$ implies $\theta_1 = \theta_2$. Using the result given by Zehna eliminates the necessity of verifying that $g$ is one-to-one. In terms of the Goodman-Kruskal (henceforth, G-K) model, this means that substituting MLE's of $P_{11}$, $P_{10}$, and $P_{01}$ in (1) yields a MLE of $x_1$, $x_2$, $k$, and $r$, respectively. Here, for example, $r$ is a function that maps $P_{11}$, $P_{10}$ and $P_{01}$ (a point in a 3-space) to a point on the real line. It is known (Johnson and Kotz, 1969, p. 288) that if the observed frequencies $a$, $b$, $c$, and $d$ have a multinomial distribution, then $a/n$, $b/n$, $c/n$ and $d/n$ are MLE's of $P_{11}$, $P_{10}$, $P_{01}$ and $P_{00}$, respectively. Applying Zehna's theorem, we have as MLE's

$$\left\{\begin{aligned} \hat{x}_1 &= \tfrac{1}{2} \mp \frac{2n^{-1}(a+b)-1}{2(1-4\hat{r})^{1/2}} \\ \hat{x}_2 &= \tfrac{1}{2} \mp \frac{2n^{-1}(a+c)-1}{2(1-4\hat{r})^{1/2}} \\ \hat{k} &= \tfrac{1}{2}(1 \pm (1-4\hat{r})^{1/2}) \\ \hat{r} &= \frac{an^{-1}-(a+b)(a+c)n^{-2}}{1-2(b/n+c/n)} \end{aligned}\right. \tag{4}$$

Another desirable property of $\hat{x}_1$, $\hat{x}_2$ and $\hat{k}$ is that they are consistent (Wilcox, 1976). A statistic $\hat{\theta}$ is a consistent estimator of a parameter $\theta$ if the probability that $\hat{\theta}$ differs from $\theta$ by more than some positive real number, say $\varepsilon$, approaches zero as $n$, the sample size, gets large. The point is that $\varepsilon$ can be chosen arbitrarily close to zero implying that $\hat{\theta}$ can be made to be arbitrarily close to $\theta$ by choosing $n$ sufficiently large. While a consistent estimator is desirable, consistency is not considered to be a strong property. This is particularly true when $\hat{\theta}$ is a MLE of $\theta$ since a MLE is nearly always consistent. However, there are several known examples in which a MLE is not consistent; this is our motivation for mentioning that $\hat{x}_1$, $\hat{x}_2$, and $\hat{k}$ have this property.

It should also be mentioned that $\hat{x}_1$, $\hat{x}_2$, and $\hat{k}$ are probably biased statistics, i.e., $E(\hat{x}_1) \neq x_1$, $E(\hat{x}_2) \neq x_2$ and $E(\hat{k}) \neq k$ where $E$ denotes expectation. The reason is that if $E(\hat{\theta}) = \theta$, then usually $E(g(\hat{\theta})) \neq g(\theta)$ for a function $g$. While an unbiased statistic is frequently (but not always) considered to be more desirable than a biased one, the use of biased MLE's is a common occurence and can give good results. Perhaps the best known example of a biased MLE is the usual Pearson product-moment correlation coefficient which may be used to estimate the correlation for a normal bivariate distribution.

A comment concerning $\hat{x}_1$, $\hat{x}_2$ and $\hat{k}$ needs to be made. The first point is that $\hat{x}_1$, $\hat{x}_2$ and $\hat{k}$ are MLE's of $x_1$, $x_2$ and $k$ assuming that $0 \leq P_{ij} \leq 1$ $(i = 0, 1; j = 0, 1)$ and that $P_{11} + P_{10} + P_{01} + P_{00} = 1$. The point is that we do not make the assumption that the G-K model holds; we merely assume that the multinomial distribution is appropriate in terms of the sampling plan used to select the examinees. The second point is that the value of $\hat{r}$ may be greater than ¼. If we accept an estimate of $r$ that is greater than ¼ then we have in effect rejected the G-K model since the resulting estimate of $k$, a complex number, cannot be interpreted as a proportion.

Before continuing, it is worth mentioning that the statistics $\hat{x}_1$, $\hat{x}_2$, and $\hat{k}$ are MLE's of $x_1$, $x_2$, and $k$ under the assumption that the G-K model holds and that they have admissible values. For example, if we assume the model holds, then $k$ must have a value between zero and one for the model to make sense. The difficulty is that with probability greater than zero $k$ can have a complex and hence an inadmissible value even when the model holds. Our early experiences with the G-K model indicate that inadmissible estimates of $x_1$, $x_2$, and $k$ occur as a result of finding $(\hat{P}_{1.} - \frac{1}{2})(\hat{P}_{.1} - \frac{1}{2}) < 0$, where $\hat{P}_{1.} = (a + b)/n$ and $\hat{P}_{.1} = (a + c)/n$ are MLE's of $P_{1.}$ and $P_{.1}$. If we then accept $\hat{P}_{1.}$ and $\hat{P}_{.1}$ as reasonable values of $P_{1.}$ and $P_{.1}$, we have rejected the G-K model since the conditions given by (3) do not hold. Yet there may be circumstances in which one is willing to assume the model holds despite any evidence that $(P_{1.} - \frac{1}{2})(P_{.1} - \frac{1}{2})$ is less than zero. A partial solution to this problem can be obtained if one is willing to assume that $P_{1.}$ and $P_{.1}$ are both greater than $\frac{1}{2}$. If, for example, we assume that $P_{1.} > \frac{1}{2}$, then $\hat{P}_{1.}$ is not a MLE of $P_{1.}$. To obtain a MLE of $P_{1.}$ we would use

$$\hat{P}'_{1.} = \begin{cases} (a + b)/n & \text{if } (a + b)/n > \frac{1}{2} \\ \frac{1}{2} & \text{if } (a + b)/n \leqslant \frac{1}{2} \end{cases}$$

Similarly, we would estimate $P_{.1}$ with

$$\hat{P}'_{.1} = \begin{cases} (a + c)/n & \text{if } (a + c)/n > \frac{1}{2} \\ \frac{1}{2} & \text{if } (a + c)/n \leqslant \frac{1}{2} \end{cases}$$

*Estimating the probability of a similar response.* What is an appropriate measure of association based on the model described above? One possibility is to use $x_1x_2 + (1 - x_1)(1 - x_2) = S$, say, the probability that a randomly chosen person answers the two questions similarly (Goodman and Kruskal, 1959, p. 151). This probability may prove to be useful in certain situations. In terms of $P_{1.}$ and $r$, it can be written as:

$$S = \left[\frac{1}{2} + \frac{2P_{1.} - 1}{2(1 - 4r)^{1/2}}\right]\left[\frac{1}{2} + \frac{2P_{.1} - 1}{2(1 - 4r)^{1/2}}\right] + \left[\frac{1}{2} - \frac{2P_{1.} - 1}{2(1 - 4r)^{1/2}}\right]\left[\frac{1}{2} - \frac{2P_{.1} - 1}{2(1 - 4r)^{1/2}}\right]$$

Applying the theorem by Zehna, we see that $\hat{x}_1\hat{x}_x + (1 - \hat{x}_1)(1 - \hat{x}_2)$ is a maximum likelihood estimator of $S$. Recall that $0 \leqslant x_i \leqslant \frac{1}{2}$, $i = 1, 2$ and note that

$$\frac{1}{2} \leqslant x_1x_2 + (1 - x_1)(1 - x_2) \leqslant 1 \tag{5}$$

is required. Furthermore, expression (5) obtains its minimum at $x_1 = x_2 = \frac{1}{2}$ and attains its maximum at $x_1 = x_2 = 0$.

*The distribution of $\hat{k}$.* In this section we derive the probability density function of $\hat{k}$ which may be useful in making interval estimates of $k$. That is, we want to determine an expression that gives the probability that the estimate of $k$ is $\hat{k}$ where $\hat{k}$ is a particular admissible value of $k$. To accomplish this task it will be convient to make a distinction between a random variable and its values. Accordingly, we define the random variable $A$ to be the number of examinees who answer both test items correctly. We let $B$ denote the random variable corresponding to the number of examinees who are correct on item 1 but incorrect on item 2 and we let $C$ be the random variable corresponding to an incorrect-correct response. Finally, $D$ will be the random variable corresponding to the number of examinees who give an incorrect response to both items. The values of $A$, $B$, $C$, and $D$ that we observe are $a$, $b$, $c$, and $d$, respectively. The random variable corresponding to $\hat{k}$ will be denoted as $\hat{K}$ and is defined as $\hat{K} = \frac{1}{2}(1 - (1 - 4\hat{R})^{1/2})$ where $\hat{R}$ is given in (6) below. As already indicated, our goal is to determine the probability density function of $\hat{K}$, i.e., we want to find an expression in terms of $P_{11}$, $P_{10}$, $P_{01}$ and $P_{00}$ that gives the probability that the random variable $\hat{K}$ will have the value $\hat{k}$. We denote this probability as $Pr(\hat{K} = \hat{k})$. The method we use to accomplish our goal is the usual change of variable technique for discrete random variables. (See Hogg and Craig, 1970, section 4.2.) We define the following transformations:

$$\begin{cases} Z_1 = A/n \\ Z_2 = (A + B)/n \\ \hat{R} = \dfrac{A/n - ((A + B)/n)((A + C)/n)}{1 - 2((B + C)/n)} \end{cases} \tag{6}$$

These transformations will yield the probability density function of $\hat{R}$, the random variable corresponding to $\hat{r}$, which in turn can be used to derive the probability density function of $\hat{K}$.

We want to show, for reasons explained below, that the three transformations defined above are one-to-one. This means that we want to show that from (6) we can obtain an expression for $A$, $B$, and $C$ in terms of $Z_1$, $Z_2$, and $\hat{R}$. From expression (6) we have that $A = nZ_1$. Substituting $nZ_1$ for $A$ gives $Z_2 = (nZ_1 + B)n$. Thus, $B = n(Z_2 - Z_1)$. Substituting $nZ_1$ for $A$ and $n(Z_2 - Z_1)$ for $B$ gives

$$\hat{R} = \frac{Z_1 - Z_2(Z_1 + C/n)}{1 - 2(Z_2 - Z_1 + C/n)} \tag{7}$$

Thus, $\hat{R} - 2\hat{R}(Z_2 - Z_1) - 2\hat{R}C/n = Z_1 - Z_2Z_1 - CZ_2/n$. Hence, $(Z_2 - 2\hat{R})C/n = 2\hat{R}(Z_2 - Z_1) - \hat{R} + Z_1 - Z_2Z_1$. Whence, $C = (n/(Z_2 - 2\hat{R}))(Z_1 - Z_2Z_1 - \hat{R}(1 - 2(Z_2 - Z_1)))$. This proves that the transformation defined by the expressions in (6) is one-to-one.

As indicated earlier, the event that the random variable $A$ has the value $a$ is written as $A = a$. Correspondingly, $Pr(A = a)$ denotes the probability that the random variable $A$ has the value $a$. Extending this notation further, the probability of the event $A = a$, $B = b$, and $C = c$ is denoted as $Pr(A = a,\ B = b,\ C = c)$ and is equal to

$$\frac{n!P_{11}^{a}P_{10}^{b}P_{01}^{c}(1 - P_{11} - P_{10} - P_{01})^{n-a-b-c}}{a!b!c!(n - a - b - c)!} \tag{8}$$

By showing that the transformation given by (6) is one-to-one, we have established that the event $Z_1 = z_1$, $Z_2 = z_2$ and $\hat{R} = \hat{r}$ corresponds to one and only one event $A = a$, $B = b$, $C = c$. Consequently,

$$Pr(Z_1 = z_1,\ Z_2 = z_2,\ \hat{R} = \hat{r}) = Pr(A = a,\ B = b,\ C = c)$$

for some unique set of values $a$, $b$, and $c$—the values being determined by (6). For example, if $n = 10$ and $a = b = c = 0$, then $z_1 = z_2 = \hat{r} = 0$. Conversely, if $z_1 = z_2 = \hat{r} = 0$, then $a = b = c = 0$. Consequently, the event $A = 0$, $B = 0$ and $C = 0$ is equivalent to $Z_1 = 0$, $Z_2 = 0$ and $\hat{R} = 0$ which implies that $Pr(A = 0,\ B = 0,\ C = 0) = Pr(Z_1 = 0,\ Z_2 = 0,\ \hat{R} = 0)$. Thus, $Pr(Z_1 = 0,\ Z_2 = 0,\ R = 0)$ can be evaluated by substituting values of zero for $a$, $b$, and $c$ in expression (8) to obtain $(1 - P_{11} - P_{10} - P_{01})^{10}$ since $n = 10$. More generally if we substitute in (8) $nz_1$ for $a$, $n(z_2 - z_1)$ for $b$ and

$$(n/(-2\hat{r} + z_2))[z_1 - z_1z_2 - \hat{r}(1 - 2(z_2 - z_1))]$$

for $c$, we have the joint *p.d.f.* of $Z_1$, $Z_2$ and $\hat{R}$. Consequently, the *p.d.f.* of $\hat{R}$ is $\Sigma_{z_1} \Sigma_{z_2} Pr(Z_1 = z_1,\ Z_2 = z_2,\ \hat{R} = \hat{r})$ where the sum is taken over all possible values of $Z_1$ and $Z_2$. Be certain to notice that the possible values of $Z_1$ depend on $Z_2$ and that the possible values of $Z_2$ depend on the value of $\hat{R}$.

Now, for the transformation $\hat{K} = \frac{1}{2}(1 - (1 - 4\hat{R})^{1/2})$ we have that $\hat{R} = ((2\hat{K} - 1)^2 - 1)/-4$. That is, the probability that the random variable $\hat{K}$ has the value $\hat{k}$ is equal to the probability that $\hat{R}$ has the value $\hat{r}$ where $\hat{r}$ is uniquely determined by $\hat{k}$. Thus, substituting $((2\hat{k} - 1)^2 - 1)/-4$ in the probability density function of $\hat{R}$, we have the probability density function of $\hat{K}$. We are aware that for practical purposes, the exact probability density function of $\hat{K}$ will be very difficult to use. A partial solution to this problem is to determine the asymptotic distribution of $\hat{K}$ which we do in a subsequent section.

*Distribution of* $\hat{X}_1$ and $\hat{X}_2$. Without loss of generality, assume that the minus sign is chosen in the expressions for $\hat{X}_1$ and $\hat{X}_2$. Let the transformations $Z_1$ and $Z_2$ be defined as in the previous section and consider

$$\hat{X}_1 = \frac{1}{2} - \frac{(2/n)(A + B) - 1}{2(1 - 4\hat{R})^{1/2}} \tag{9}$$

We want to show that the transformation is one-to-one. To do this, it is sufficient to find a unique solution for $C$ in terms of $Z_1$, $Z_2$ and $\hat{X}_1$. As before, $A = nZ_1$ and $B = n(Z_2 - Z_1)$. From (9), $2(\hat{X}_1 - \frac{1}{2}) = -((2/n)(A + B) - 1)/(1 - 4\hat{R})^{1/2}$. Hence,

$$1 - 4\hat{R} = \left[\frac{(2/n)(A + B) - 1}{2(\hat{X}_1 - \frac{1}{2})}\right]^2.$$

Solving for $\hat{R}$ we have

$$\hat{R} = \tfrac{1}{4}\left[1 - \frac{(2(A+B)/n) - 1}{2\hat{X}_1 - \tfrac{1}{2}}\right]^2 .$$

Substituting $nZ_1$ for $A$ and $n(Z_2 - Z_1)$ for $B$ we find that

$$\hat{R} = \tfrac{1}{4}\left[1 - \left(\frac{(2Z_2) - 1}{2\hat{X}_1 - \tfrac{1}{2}}\right)^2\right] \tag{10}$$

Substituting the right hand side of (10) for $\hat{R}$ in (7) gives a solution for $C$ in terms of $Z_1$, $Z_2$, and $X_1$. Thus, if we substitute in (8) $nz_1$ for $a$, $n(z_2 - z_1)$ for $b$ and $(n/(-2r + z_2))[z_1 - z_2z_1 - \hat{r}(1 - 2(z_2 - z_1))]$ for $c$. where

$$\hat{r} = \frac{1 - [(2z_2 - 1)/(2\hat{x}_1 - \tfrac{1}{2})]^2}{4},$$

we have the joing *p.d.f.* of $Z_1$, $Z_2$ and $\hat{X}_1$. Then $\Sigma_{z_1} \Sigma_{z_2} Pr(z_1, z_2, \hat{x}_1)$ gives the *p.d.f.* of $\hat{X}_1$ where the sums are taken over all possible values of $Z_1$ and $Z_2$. The *p.d.f.* of $\hat{X}_2$ can be derived in a similar fashion.

*The asymptotic distribution of $\hat{X}_1$, $\hat{X}_2$, and $\hat{R}$.* The exact probability density functions of $\hat{X}_1$, $\hat{X}_2$ and $\hat{R}$ are not in a form that is easy to use. For example, the probability density function of $\hat{X}_1$ as given above involves summing over two variables, the admissible values of which depend on the value of $\hat{X}_1$. Therefore, we show that, as the number of observations gets large, the distribution of $\hat{X}_1$, $\hat{X}_2$ and $\hat{R}$ approaches normality. We use the following theorem given by Wilks (1962, section 9.3): Suppose $V_{1m}, \ldots, V_{jm}$, $j = 1, \ldots, n$, is a sample from a $j$-dimensional distribution with finite means, say $\mu_j$, and positive definite variance-covariance matrix $\|\sigma_{im}\|$ $(i, m = 1, \ldots, j)$. Let $g(V_1, \ldots, V_j)$ be a function and suppose the first derivatives of $g$, say $\partial g/\partial V_i = g_i$, $i = 1, \ldots, j$ exist at all points in some neighborhood of $(\mu_1, \ldots, \mu_j)$ and let $g_i^\circ = g(\mu_1, \ldots, \mu_j)$. Then, if at least one of the $g_i^\circ$ is not equal to zero, $g(\bar{V}_1, \ldots, \bar{V}_j)$ is asymptotically normal with mean $g(\mu_1, \ldots, \mu_j)$ and variance $(1/n)\Sigma_{i,m} g_i^\circ g_m^\circ$ where $\bar{V}_i = (1/n)\Sigma_{m=1}^{n} V_{im}$.

We verify the conditions of the theorem for $\hat{R}$. The proof of the asymptotic normality of $\hat{X}_1$ and $\hat{X}_2$ is virtually the same as the proof for $\hat{R}$ and is therefore omitted. For $\hat{R}$, we may set $\bar{V}_1 = A/n$, $\bar{V}_2 = B/n$, and $\bar{V}_3 = C/n$ since $A/n$, $B/n$, and $C/n$ are the sample means for estimating $P_{11}$, $P_{10}$, and $P_{01}$, respectively. To verify that the variance-covariance of $A$, $B$, and $C$ is positive definite it is sufficient to observe that it is impossible to find constants $c_1$, $c_2$, $c_3$, $c_4$ such that

$$Pr(c_1A + c_2B + c_3C = c_4) = 1$$

(Wilks, 1962, section 3.5.1). We must assume, however, that $P_{00} < 1$.

Finally, we omit the tedious algebra and merely state that, in general, $g_i^\circ \neq 0$ and exists for some neighborhood of $P_{11}$, $P_{10}$ and $P_{01}$. However, it is necessary to assume that $r \neq 1/4$ and that $P_{10} + P_{01} \neq 1/2$ since otherwise $g_i^\circ$ does not exist due to division by zero.

*Modification one of the Goodman-Kruskal model.* In this section we describe the first of two modifications of the Goodman-Kruskal model. We bring about this modification by altering the definition of $x_1$ and $x_2$. In particular, we assume that the probability of not knowing either item 1 or item 2 yet getting it right is zero, and so we define $x_1$ as the probability of getting item 1 wrong given that the examinee knows the answer. This assumption appears to be appropriate when guessing is essentially ruled out by the nature of the item, which can be true for completion items. The parameter $x_2$ denotes the comparable probability for item 2. The parameter $k$ retains its earlier meaning. The resulting fourfold table of probabilities is given in Table 2.

**Table 2**

| | | Item 2 | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Item 1 | 1 | $k(1 - x_1)(1 - x_2)$ | $k(1 - x_1)x_2$ | $k(1 - x_1)$ |
| | 0 | $kx_1(1 - x_2)$ | $kx_1x_2 + 1 - k$ | $kx_1 + 1 - k$ |
| | | $k(1 - x_2)$ | $kx_2 + 1 - k$ | 1 |

Solving for $x_1$, $x_2$ and $k$ in terms of $P_{ij}(i, j = 0, 1)$ we have: $x_1 = P_{01}/P_{.1}$, $x_2 = 1 - (P_{11}/P_{1.}) = P_{10}/P_{1.}$, $k = P_{.1}P_{1.}/P_{11}$, where $P_{1.} = P_{11} + P_{10}$ and $P_{.1} = P_{11} + P_{01}$. Notice that $P_{.1}P_{1.}$ must be less than $P_{11}$ for the model to hold since otherwise $k$, a proportion, is greater than one. Applying the theorem by Zehna we have that $\hat{x}_1 = (c/n)/((a + c)/n) = c/(a + c)$, $\hat{x}_2 = 1 - (a/(a + b)) = b/(a + b)$, $\hat{k} = ((a + c)/a) \cdot ((a + b)/n)$ are MLE's of $x_1$, $x_2$ and $k$, respectively, where $a$, $b$, $c$, and $d$ are the observed frequencies.

We do not give the exact probability density functions of $\hat{x}_1$, $\hat{x}_2$ or $\hat{k}$ since, at this time, we believe that such information would be of little practical use. It should be mentioned, however, that $\hat{x}_1$, $\hat{x}_2$ and $\hat{k}$ are all asymptotically normally distributed.

*Modification two of the Goodman-Kruskal model.* In the second modification of the Goodman-Kruskal model, we assume that the probability of getting an item wrong given that the examinee knows the answer is zero. In contrast to the first modification, we assume that the probability of not knowing the answer to item 1 yet getting it right is greater than zero, and so we define $x_1$ as the probability of getting item 1 right given that the examinee does not know the answer. The parameter $x_2$ denotes the comparable probability of item 2. The fourfold table of probabilities is:

| | | Item 2 | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Item 1 | 1 | $k + (1 - k)x_1x_2$ | $(1 - k)x_1(1 - x_2)$ | $k + (1 - k)x_1$ |
| | 0 | $(1 - k)(1 - x_1)x_2$ | $(1 - k)(1 - x_1)$<br>$\cdot (1 - x_2)$ | $(1 - k)(1 - x_1)$ |
| | | $k + (1 - k)x_2$ | $(1 - k)(1 - x_2)$ | 1 |

Solving for $x_1$, $x_2$ and $k$ we have $x_1 = P_{10}/P_{.0}$, $x_2 = P_{01}/P_{0.}$, $k = 1 - (P_{.0}P_{0.}/P_{00})$ where $P_{.0} = P_{10} + P_{00}$ and $P_{0.} = P_{01} + P_{00}$. The theorem by Zehna implies that $\hat{x}_1 = b/(b + d)$, $\hat{x}_2 = c/(d + c)$, $\hat{k} = 1 - ((c + d)/d)((b + d)/n)$ are MLE's of $x_1$, $x_2$, and $k$ respectively. As was the case for the two earlier models, $\hat{X}_1$, $\hat{X}_2$, and $\hat{K}$ are consistent, probably biased, and asymptotically normal.

*Note on estimation.* Preliminary trials of the methods of estimating $x_1$, $x_2$, and $k$ have been made with data for algebra items administered to high school students. In a number of instances inadmissible values were secured for the estimate assuming the Goodman-Kruskal model. These inadmissible values appeared to be a function of the instructional history of the examinees, and/or the assumptions of the particular model used (modification one of the Goodman-Kruskal model appeared to be more appropriate for the test items used), and/or of sampling fluctuations in the data.

Another estimation method known as STEPIT (Chandler, 1969) estimates a set of parameter values which fit a model to data, with the model not required to be linear or approximately linear in the parameters. STEPIT was employed for several of the data sets (two-by-two tables) for which inadmissible values had been secured from the MLE's, and gave real valued, reasonable estimates. STEPIT was also employed for several data sets for which the MLE's were reasonable, and it was found that the STEPIT estimates were essentially identical to the MLE's. Apparently the analytic solutions, when they are admissible, are "correct" in that an *ad hoc* procedure which capitalizes on the idiosyncracies of the particular set of data gives essentially the same solution. It also seems likely that this STEPIT procedure can give admissible solutions when the analytic one is unsatisfactory.

### Table of Symbols

$P_{11}$ = The probability that a randomly chosen examinee answers both test items correctly.

$P_{10}$ = The probability that a randomly chosen examinee gives a correct-incorrect response.

$P_{01}$ = The probability that a randomly chosen examinee gives an incorrect-correct response.

$P_{00}$ = The probability that a randomly chosen examinee gives an incorrect response to both test items.

$k$ = The population parameter denoting the proportion of examinees who know the answer to both test items.

$A$ = The random variable that corresponds to the number of examinees who answer both test items correctly.

$a$ = The value of the random variable $A$ which is observed for a particular sample of examinees. That is, $a$ is the number of examinees who answer the test item correctly on both occasions.

$B$ = The random variable corresponding to the number of examinees giving a correct-incorrect response to the test item pair.

$b$ = The value of the random variable $B$. Equivalently, $b$ is the observed number of examinees giving a correct-incorrect response for a particular sample of examinees.

$C$ = The random variable corresponding to the number of examinees who give an incorrect-correct response.

$c$ = The observed value of the random variable $C$.

$D$ = The random variable corresponding to the number of examinees who give an incorrect-incorrect response.

$d$ = The observed value of the random variable $D$.

$x_1$ = The population parameter denoting the probability of an inappropriate response to test item one. The definition of an inappropriate response depends on which model is assumed.

$x_2$ = The population parameter denoting the probability of an inappropriate response to item two. The definition of an inappropriate response depends on which model is assumed.

$$r = \frac{P_{11} - P_{1.}P_{.1}}{1 - 2(P_{12} + P_{01})}$$

$\theta$ = An arbitrary population parameter.

$$\hat{R} = \frac{(A/n) - ((A + B)/n)((A + C)/n)}{1 - 2((B + C)/n)}$$

The symbol "^" was used in order to make a distinction between the value of $\hat{R}$, $\hat{r}$, and the parameter.

$\hat{r}$ = The observed value of the random variable $\hat{R}$. It is also a MLE of $r$.

$\hat{K} = \frac{1}{2}(1 - (1 - 4\hat{R})^{1/2})$

$\hat{k}$ = The observed value of $\hat{K}$ that is used to estimate $k$.

$$\hat{X}_1 = \frac{1}{2} - \frac{(2(A + B)/n) - 1}{2(1 - 4\hat{R})^{1/2}}$$

$\hat{x}_1$ = The observed value of the random variable $\hat{X}_1$ that is used to estimate the parameter $x_1$.

## Chapter 6

# ABSTRACTS OF SELECTED JOURNAL ARTICLES

**Andrea Pastorok Pearlman**

Articles judged to be related to the material presented in this monograph were abstracted from the following journals:

American Educational Research Journal, 1968 through 1975
Educational and Psychological Measurement, 1968 through 1975
Journal of Educational Measurement, 1964 through 1975

The articles abstracted here do not appear in the list of references cited, nor in the list of additional references. The list of additional references is primarily a selected list of references cited by the authors of the articles in this section.

Anderson, T. H. Cloze measures as indices of achievement comprehension when learning from extended prose. *Journal of Educational Measurement*, 1974, *11*, 83-92.

Two experiments were conducted to test the hypothesis that *cloze measures* are a function of content achievement among adult learners. Given that this proposition is true, cloze measures should be sensitive to instructional treatments and should respond in a pattern similar to other indices of achievement.

The following five achievement tests were constructed and administered to college juniors and seniors:

1) a 20 item multiple-choice test
2) a reproduction passage cloze test
3) a recognition passage cloze test
4) a reproduction summary cloze test
5) a recognition summary cloze test

All tests showed significant differences between pre-and post conditions, and between recognition and reproduction modes. These results lend strong support to the proposition that cloze measures can serve as indices of achievement.

As indicated by a $\omega^2$ statistic, the reproduction summary cloze test was found to be the most sensitive to instructional treatment on pre-post measures, and the reproduction passage cloze was found to be the least sensitive.

An important finding was the failure of the summary cloze to decrease significantly over the delay interval, while the cloze scores on the passage and on the multiple-choice test decreased to such a level that they were not significantly different from preinstruction scores. This suggests that the outline of the set of instructional materials was well retained after a month's delay, even though many of the details were not retrievable.

Anderson concludes that cloze measures are psychometrically similar to those from a good multiple-choice test when assessing pre and post instruction differences in achievement comprehension. This suggests that equivalent forms of comprehension achievement tests can be constructed by the cloze method.

**Baker, E. L. The effects of manipulated item-writing constraints on the homogeneity of test items. *Journal of Educational Measurement*, 1971, *8*, 305-309.**

The purpose of this study was to determine the effect of using several options on the relative homogeneity of the produced test items. The four conditions for writing test items were:
a) general objective
b) behavioral objective
c) behavioral objective plus test item
d) behavioral objective plus item-form
Two tests, one in current events and one in subtraction, were constructed by selecting four items generated from each of the four experimental conditions. These tests were administered to 51 seventh graders in Los Angeles. The expected tendencies toward greater homogeneity among items produced under the three conditions employing behavioral objectives were not found in this study. Homogeneity was measured in terms of interitem correlations (phi coefficients) which were averaged using the $r$ to $z$ transformation.

Baker, F. B. Origins of the item parameters $X_{50}$ and $\beta$ as a modern item analysis technique. *Journal of Educational Measurement*, 1965, *2*, 167-180.

The purpose of this paper is to bring together the developments relevant to the curve fitting methods of item analysis. The approach is to present the developments in essentially chronological order from its inception in the Binet studies to its modern implementation on digital computers.

The author points out that the modern digital computer has freed us from nearly all constraints due to data processing or computation associated with item analysis, therefore we should not continue to operate under yesterday's limitations. He notes that despite Lawley's paper showing that mental test theory should begin with the specification of the characteristics of the items within an instrument and that subsequent theory should be built upon the item parameters, most of the current mental test theory begins with the test score and ignores the underlying composition of that score. Baker concludes that full advantage of the technological advances can be made only when modern item analysis techniques become an integral part of the total process of test development.

Barcikowski, R. S. The effects of item discrimination on the standard errors of estimate associated with item-examinee sampling procedures. *Educational and Psychological Measurement*, 1974, *34*, 231-237.

A Monte Carlo study was conducted using item-examinee sampling procedures to examine the standard error of estimate for a given test's mean and variance. The main variables considered were test length, item difficulty, and item discrimination. The results indicate that optimal estimates, i.e., smallest standard error, of both mean and variance from a single item-examinee sampling plan may not be possible.

Barcikowski, R. S. A Monte Carlo study of item sampling (versus traditional sampling) for norm construction. *Journal of Educational Measurement*, 1972, *9*, 209-214.

Using a computer-based model of an item trace line, a random sampling experiment concerned with comparing item sampling estimates to traditional (examinee) sample estimates of the mean and variance of the distribution of test scores was conducted.

Item sampling and traditional sampling were studied with large numbers of simulated subjects across several different types of tests (e.g., tests having different combinations of item difficulties and biserial correlations). The results indicate that the optimal method for estimating a test's parameters may depend on several conditions. Under all conditions, item sampling proved superior to traditional sampling in estimating population test means. However, with certain test lengths, ranges of item difficulty, and discrimination, traditional sampling provided better estimates of test variance than did item sampling.

These results indicate that in deciding on the data-gathering design to be used in seeking norm information, attention should be given to item characteristics and test length with particular attention paid to the range of biserial correlations between item response and ability.

Baskin, D. A configuration-scoring paradigm for identical raw scores. *Journal of Educational Measurement*, 1975, *12*, 3-5.

Traditional test-scoring does not allow the examination of differences among subjects obtaining identical raw scores on the same test. The author develops and illustrates a configuration-scoring paradigm which minimizes the number of digits needed to report configuration-scores, while simultaneously providing a numerical basis upon which to compare characteristics of subjects having identical raw scores.

Beck, M. D. Achievement test reliability as a function of pupil-response procedures. *Journal of Educational Measurement*, 1974, *11*, 109-114.

The present study was designed in part to assess the differential effect of two pupil response procedures on *Metropolitan Achievement Tests* scores of third and fourth grade pupils. Results indicate that the reliability of scores is not significantly altered when pupils respond to achievement test items on separate answer folders rather than directly in their test booklets.

Board, C. & Whitney, D. R. The effect of selected poor item-writing practices on test difficulty, reliability and validity. *Journal of Educational Measurement*, 1972, *9*, 225-233.

The major purpose of this study was to investigate the effect of selected poor item-writing practices on test difficulty, reliability and validity.

Within the limitations of this study, the authors believe that the following conclusions are warranted:

1) "Window dressing" or extraneous material in item stems makes the test items easier for poor students but more difficult for better students. There is little overall effect on test difficulty. Presence of this flaw reduces the internal consistency of the test.

2) Incomplete stems make the test items more difficult for most students. Presence of this flaw reduces the internal consistency of the test.

3) Using a keyed response which differs in length from the distractors does not make test items less difficult. Poor students gain more from this flaw, however, than do good students. Presence of this flaw reduces both the internal consistency and validity of the test.

4) Grammatical consistency between stem and keyed response does not have a major effect on test difficulty. Presence of this flaw, however, reduces the validity of the test.

The authors recommend that future studies of item-writing practices incorporate achievement as a blocking or control variable because of the presence of interaction effects in this study. At least for the principles studies here, the authors contend that poor item-writing practices serve to obscure (or attenuate) differences between good and poor students—chiefly by making the latter look more like the former than their scores on "error-free" tests would suggest.

Bormuth, J. R. Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, 1968, *5*, 189-196.

The purpose of this study was to establish a set of criterion scores for cloze readability tests which would be comparable to the criterion scores used with oral reading tests employed to measure the readability of passages. A secondary purpose was to determine the extent of correlation between passage difficulties determined using cloze tests and those determined using comprehension and word recognition tests.

The following conclusions were made:

1) The cloze scores comparable to the comprehension criterion scores of 75% and 90% were about 44% and 57%, respectively, on the tests used in this study. These cloze scores probably do not differ greatly from those that would have been obtained had the comprehension tests been written by another author following the same item-writing rules.

2) The cloze scores comparable to the word recognition criterion scores of 95% and 98% were about 33% and 54%, respectively.

3) There were large differences between the cloze criterion scores obtained when comprehension scores were used as the criterion and those obtained when word recognition scores were used as the criterion. This constitutes grounds for suspecting that, contrary to tradition, the word recognition and comprehension criterion scores are not comparable.

4) Cloze tests seem to be highly valid measures of passage difficulty. Passage difficulties determined using cloze tests exhibited correlations ranging from .90 to .96 with passage difficulties determined using comprehension and word recognition tests.

Bowers, J. A note on Gaylord's "Estimating test reliability from the test-item correlations." *Educational and Psychological Measurement*, 1971, *31*, 427-429.

The author examined the algebraic consistency of Guilford's reliability formula which Gaylord (*Educational and Psychological Measurement*, 1969, *29*, 303-304) demonstrated. Guilford's formula is shown to be erroneous. The author noted that Kuder-Richardson reliabilities depend on item inter-correlations.

Brandenburg, D. C. & Whitney, D. R. Matched pair true-false scoring: Effect on reliability and validity. *Journal of Educational Measurement*, 1972, *9*, 297-302.

The matched pair technique for writing and scoring true-false (T-F) items was designed to compensate for the acquiescence response (responding affirmatively when in doubt about an answer) of primary school children. The Primary Test of Economic Understanding (PTEU) was designed to be scored using the matched pair procedure.

Five scoring methods were used:

1) *Traditional*—one point was given for each correct response.

2) *Matched pair scoring*—items were written in pairs of one T and one F on the same concept. Credit was given only if the student answered both correctly.

3) *Random matched pairs*—one item was selected at random from the T items and one item was selected at random from the F items.

4) *Modified matched pair scoring*—two points were given for correctly answering both members of an item pair (like Method 2) but also one point was awarded for answering the false item correctly.

5) *Differential credit* was given for correct T and F (4/3 point for correct false; 2/3 point for correct T).

*Results*

Alpha coefficients for each of the five methods were computed. Median alphas suggest that Methods 2 and 4 were more internally consistent than traditional scoring.

To test for concurrent validity, product moment correlations between scores arising from each scoring method and selected Iowa Tests of Basic Skills (ITBS) standard scores were computed. Contrary to expectation, the correlations of traditional and matched pair scores with ITBS subtests (when adjusted for differing reliabilities) were approximately equal.

The authors conclude that although matched pair scoring does offer a way of increasing internal consistency of scores arising from T-F tests, it seems unlikely that the gains were accomplished by reducing the acquiescence set.

Brennan, R. L. A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, 1972, *32*, 289-303.

In the first section of this paper the author discussed a rationale for upper-lower types of discrimination indices and the relationship between this rationale and the discrimination index D. Although Brennan considers the D index to be useful, he notes that the necessity for using equal numbers of observations in the upper and lower groups seems overly restrictive. He points out that in the case of mastery tests, criterion-referenced tests and many teacher-made tests, the expectation is that most of the students will get most of the items correct yielding a distribution of test scores that is negatively skewed. Brennan developed a new upper-lower discrimination index called B that allows for the use of unequal $n$'s in the upper and lower groups, thereby giving the evaluator the freedom to choose appropriate cut-off points between these groups. The exact distribution of the index is determined under the null hypothesis $B = 0$.

Brennan states that for norm-referenced tests, nondiscriminating items and negatively discriminating items are usually unacceptable, while positively discriminating items are acceptable. For criterion-referenced tests, the interpretation of discrimination indices needs to be modified. According to Brennan, the ideal item in the criterion-referenced testing situation is the item with a nonsignificant discrimination index and a high difficulty level; items that discriminate negatively are clearly unacceptable; and items that discriminate positively usually indicate a need for revision.

Brennan, R. L. The calculation of reliability from a split-plot factorial design. *Educational and Psychological Measurement*, 1975, *35*, 779-788.

This paper treats the question, "How should one estimate the reliability of schools (or classrooms)?" The author reviews the use of variance components in the estimation of reliability (or generalizability) coefficients in a split-plot factorial design (SPF) with persons nested within schools.

Through the use of variance components from the SPF design, he derives estimates of reliability for schools and for persons within schools. He then compares the reliability for persons within schools from a SPF design with the reliability for persons from a randomized block design. Finally, he compares the reliability for schools from a SPF design with the reliability for school means from a randomized block design.

Burnett, J. D. Parallel Measurements and the Spearman-Brown formula. *Educational and Psychological Measurement*, 1974, *34*, 785-788.

The author reviews the general use of the Spearman-Brown formula for calculating the reliability of parallel tests with different lengths. Three uses of it are presented:

1) to calculate the reliability of a lengthened or shortened test,

2) to facilitate comparison among parallel tests of different lengths with different reliabilities by converting all the tests to an hypothesized length,

3) to calculate how many items are necessary to add to a test to raise the test's reliability to a specified level.

The author emphasizes the necessity of meeting the assumption that the component tests be parallel. The property that the parallel tests be non-negatively correlated is derived. He concludes that one should pay close attention to the theory underlying an analysis.

Carver, R. P. A model for using the final examination as a measure of the amount learned in classroom learning. *Journal of Educational Measurement*, 1969, *6*, 59-68.

A student's score on the final examination in a classroom learning situation does not necessarily represent the amount learned during the course. Various measures of gain have been advanced to measure the amount learned, but all have subsequently been found inadequate. It is hypothesized that the relationship between test scores and knowledge is curvilinear. A rationale is presented for the curvilinear nature of the posited relationship and for the fit of the model to classroom learning. From hypothetical data conforming to the model expressed in a mathematical formula, it was shown that it is possible for the final examination to be the best indicant of amount learned, even though individuals are not equal in proficiency at the beginning of the learning task. Based upon several considerations it was concluded that, at present, the best indicant of amount learned in many classroom situations is the final examination.

Carver, R. P. Analysis of "chunked" test items as measures of reading and listening comprehension. *Journal of Educational Measurement*, 1970, *7*, 141-150.

The "chunked" type of test item was developed which required S's to recognize groups of words whose meaning had been changed from that in the original reading or listening passage. The "chunked" type of item requires the deletion of groups of words from the original passage whereas the cloze item deletes simple words.

Results of the first study indicate that the three comprehension measures—multiple-choice comprehension, chunked comprehension, and chunked accuracy—correlated approximately equally with the various tests of intelligence, aptitude, listening and reading. Individual differences on the chunked reading test were found to correlate .68 with a multiple-choice alternate form.

In a second study, data indicate that both "chunked" tests and multiple choice tests are sensitive to within individual decrements in comprehension with an increase in the speed of speech.

Results from both studies are cited to provide evidence for the validity of the chunked items as measures of comprehension. However, the author contends that other results suggest that the chunked items may be less dependent upon grammatical and vocabulary knowledged and more sensitive to within individual changes in comprehension as compared to the traditional multiple-choice question.

Carver, R. P. Rejoinder to Knapp's note. *Journal of Educational Measurement*, 1970, *7*, 52.

This article, a rejoinder to Knapp's note, deals with the problem of measuring gain.

Although the expression cited by Knapp ($\sqrt{n-i} - \sqrt{n-f}$) is the way to calculate amount learned within the confines of the specific mathematical formulas presented by Carver, the opinion was expressed that $f$ (final test score) would remain one of the best indicants of amount learned under other hypothesized curvilinear models.

Cleary, T. A. & Linn, R. L. A note on the relative sizes of the standard errors of two reliability estimates. *Journal of Educational Measurement*, 1969, *6*, 25-27.

Formulas for the standard error of parallel-test correlation and for the Kuder-Richardson formula 20 reliability estimate are provided. Given equal values of the two reliabilities in the population, the standard error of the Kudar-Richardson formula 20 is shown to be somewhat smaller than the standard error of a parallel-test correlation for reliability values, sample sizes, and test lengths that are usually encountered in practice.

Collett, L. S. Elimination scoring: An empirical evaluation. *Journal of Educational Measurement*, 1971, *8*, 209-214.

The purpose of this investigation was to compare experimentally the reliabilities and validities of three techniques for scoring multiple-choice tests:

a) classical (C)

b) weighted choice (W)

c) elimination (E) (the elimination score is the number of incorrect options eliminated minus ($k - 1$) for each correct answer eliminated, where $k$ is the number of options per item.)

Specific Hypotheses:

1) Reliability. The error in predicting $X_2$ from $X_1$ will be smaller under the *E* method than under *C* or *W* methods.

2) Criterion-Related Validity. The error of predicting *Y* scores from the summed *X* scores will be smaller under *E* than under *C* or *W* methods.

It was observed that both *r*'s and *SE*'s obtained in predicting $X_2$ from $X_1$, ranked *ECW* from best to worst.

It was observed that the pattern of *r*'s and *SE*'s obtained from the prediction of *Y* scores from the summed treatment scores ($X_1 + X_2$) was similar to that obtained in the reliability prediction: the treatments ranked *ECW* from best to worst in both cases. The results of the planned comparisons supported the hypothesis superiority of the *E* method for both the *E* vs. *C* comparison and for the *E* vs. *W* However, the *SE*'s for *C* and *W* were not significantly different.

Cox, R. C. Item selection techniques and evaluation of instructional objectives. *Journal of Educational Measurement*, 1965, *2*, 181-185.

The major conclusions of this study are:

1) Statistical selection of items from the total item pool has a biasing effect on the selected tests. The proportion of items in the selected tests which measure certain instructional objectives is unlike the proportion of items in the total item pool which measures the same objectives. The selected tests are not representative of the total item pool in this respect.

2) Statistical selection of items from the total item pool operates differentially for male and female groups. When the statistical data obtained from the female tryout group is used to select tests from the total item pool, the results differ from those obtained using the male tryout group. The structure of the selected tests, as indicated by the taxonomical structure of the items, differs for the male and female groups.

Cox, R. C. & Sterrett, B. G. A model for increasing the meaning of standardized test scores. *Journal of Educational Measurement*, 1970, *7*, 227-228.

The authors propose a method for obtaining criterion-referenced information from standardized tests. The model included:

a) a precise description of curriculum objectives and a definition of pupil achievement in relation to these objectives,

b) the coding of each standardized test item with reference to the curriculum,

c) the assignment of two scores to each pupil, one reflecting his achievement on items that test content to which he has been exposed, the other his achievement on items that test content beyond his present status in the curriculum or not represented in the curriculum at all.

Crawford, C. R. Item difficulty as related to the complexity of intellectual processes. *Journal of Educational Measurement*, 1968, *5*, 103-107.

Intellectual processes defined in both Bloom's (1954) taxonomy and by the Committee on Student Appraisal (1962) are considered to be hierarchical. Because of this hierarchical principle, it has been argued that items measuring the more complex processes are, by their very nature, more difficult than items measuring the less complex processes.

The purpose of this study was to investigate the relationship between item difficulty and complexity of intellectual processes presumably measured by multiple-choice items when knowledge is not held constant.

The results indicate that the order of difficulty level was, in every analysis except one, statistically different from the order of complexity. This suggests that there is not necessarily a direct relationship between the complexity of intellectual processes and the difficulty of items which purportedly measure them. This finding is consistent with Guttman's [In Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*, Glencoe, Ill.: Free Press, 1954, p. 283] statement:

> There is some danger of confusing the notion of degree of complexity with that of difficulty. If we say that subtraction is more complex than addition, we do not mean by this that subtraction is necessarily more difficult than addition. Complexity and difficulty have no necessary connection with each other in our theory.

Crehan, K. D. Item analysis for teacher-made mastery tests. *Journal of Educational Measurement*, 1974, *11*, 255-262.

The focus of this study is on item selection for teacher-made mastery tests. The author questions whether teacher-made tests resulting from various item selection techniques differ when evaluated by appropriate methods of estimating criterion-referenced reliability and validity.

Crehan adopted Carver's concept of equivalence as the reliability criterion for his study. The validity standard (which reflects the degree to which the test score discriminates between a group of "masters" and "nonmasters") was derived from item responses obtained from independent groups of instructed and non-instructed students. The author identifies passing score as the cut score which maximized estimated validity.

Six item techniques are compared.

Eighteen volunteer junior and senior high school teachers wrote behavioral objectives and parallel items for each of the original items. The entire pool of items was administered to two classes before and after instruction and to two other classes only after instruction.

Pair of tests developed by each of the six methods were derived. Estimates of test reliability and validity were obtained using responses independent of the test construction sample.

No specific selection method resulted in consistently higher reliability rankings; but the modified Brennan and Cox-Vargas methods consistently resulted in higher observed validity rankings.

The author notes that generalizations of this study are limited because of nonrandom observations. However, the author assumes that criteria employed for reliability and validity are appropriate for evaluation of teacher-made mastery tests. Crehan questioned whether the magnitude of improvement in test validity of objective item selection over teacher selection is worth the necessary effort on the part of the teacher.

Cureton, E. E. Reliability of multiple-choice tests is the proportion of variance which is true variance. *Educational and Psychological Measurement*, 1971, *31*, 827-829.

Frary (*Educational and Psychological Measurement*, 1969, *29*, 359-365) presented an analysis which seemed to show that classical weak true-score theory does not apply to multiple-choice tests. Cureton showed that the difficulty with Frary's derivation is that the guessing score is not separated into a true component and an error component.

Cureton, E. E. The stability coefficient. *Educational and Psychological Measurement*, 1971, *31*, 45-55.

The author noted that the formula he previously presented (*Educational and Psychological Measurement*, 1958, *18*, 715-738 and *Educational and Psychological Measurement*, 1965, *25*, 327-346) for the stability coefficient was essentially the same formula given by Remmers and Whistler (*Journal of Educational Psychology*, 1938, *29*, 81-92). Although the formula is correct, both his derivation and the one given by Remmers and Whistler were slightly defective. A derivation which the author believes to be more nearly correct was presented in this paper together with some further discussion.

Darlington, R. B. Some techniques for maximizing a test's validity when the criterion variable is unobserved. *Journal of Educational Measurement*, 1970, *7*, 1-14.

A set of techniques is presented for constructing a test or test battery which can be inferred to correlate as highly as possible with a hypothetical construct which is named but not measured directly. Use of the techniques requires the test constructor to describe the nature of the construct indirectly, by estimating the relative sizes of the construct's correlations with several observable variables which the test constructor has selected. Techniques are also described for estimating the validity of a test constructed by these methods.

Diamond, J. J. A preliminary study of the reliability and validity of a scoring procedure based upon confidence and partial information. *Journal of Educational Measurement*, 1975, *12*, 129-133.

This investigation concerns the estimation of the reliability and validity of scores yielded from a scoring procedure based upon confidence and partial information. The author reports that the overall result obtained from this study is that the experimental scoring procedure yields scores that, descriptively, are slightly more reliable than an inferred number-right score, but not necessarily more valid. He suggests that this experimental scoring procedure be investigated further.

Diederich, P. B. Shortcut item-test correlations for teacher-made tests. *Journal of Educational Measurement*, 1970, *7*, 43-44.

In the range of differences commonly found on teacher-made tests, the item-test correlation is approximately equal to the difference between percents correct in high and low groups, each including 27 percent of the students tested.

The author contends that with the small number of cases available to teachers, they would be well advised to round their item-test correlations to the nearest tenth. By doing this, there is no discrepancy between them and the high-low differences in percent correct throughout this range.

Ebel, R. L. Blind guessing on objective achievement tests. *Journal of Educational Measurement*, 1968, *5*, 321-325.

The purpose of this study is to determine whether or not students guessed blindly on objective tests in a course in measurement. The results seem to support the following conclusions regarding guessing on classroom tests of the kind studied, given under the conditions in this study:

1) Students do relatively little blind guessing on such tests. (The proportion of responses reported to be guesses ranged from 3% to 8%.)

2) Responses reported to be no better than blind guesses are very little better. (The percent of reported guesses that were correct ranged from 52% to 56%.)

3) Students seeking highest scores on a test are well advised to answer all questions even when the usual correction for guessing is applied. (Their blind guesses to true-false items tend to be correct more than half of the time.)

4) The distributions of guesses among students and among test items are skewed. (A few students report most of the blind guessing. A few items draw most of the blind guesses.)

5) Poor students report slightly more guessing than do students as a whole.

6) Difficult items attract considerably more guesses than do items as a whole.

Ebel, R. L. Can teachers write good true-false test items? *Journal of Educational Measurement*, 1975, *12*, 31-35.

True-false achievement test items written by classroom teachers show about two-thirds of the discrimination of their multiple-choice test items. This is about what would be expected in view of higher probability of chance success on the true-false items. However, at least half again as many true-false items as multiple choice items can be answered comfortably in the same period of time. Thus the larger number of true-false items compensates for the lower discriminating power of the individual items.

The data of this study support the belief that in the hands of typical classroom teachers, the two item forms can be expected to give approximately equal reliabilities for tests that require equal completion time.

Ebel, R. L. Confidence weighting and test reliability. *Journal of Educational Measurement*, 1965, *2*, 49-57.

The result of some hypothetical studies suggest that confidence weighting can be effective if the more capable students are also more discriminating than less capable students in choosing which responses to give confidently. But the results of recent experimental studies suggest that sometimes the more capable students are not much more successful than their less capable classmates in deciding when to answer confidently and when to answer cautiously.

It is noted that general attitudes of confidence or caution, uncorrelated or almost uncorrelated with competence, may be potent factors affecting a student's confidence weighted score on a test. To neutralize the possibly irrelevant influence of this personality trait, the author proposes to specify in advance, and identically for all students, the proportion of answers that can be, and must be, given confidently. Then, correctness of judgment as to where confidence should be placed, and correctness of the actual answers given, will be the only factors affecting the examinee's test score.

Ebel, R. L. Some measurement problems in a national assessment of educational progress. *Journal of Educational Measurement*, 1966, *3*, 11-17.

In this article, the following points are advanced by Ebel.

1) Valid educational assessment must be based on a valid conception of the nature of educational achievement.

2) In general, it is more valid to conceive of educational achievement as the acquisition of specific elements of usable knowledge than as the culmination of general mental abilities.

3) Statements of general objectives often contribute little to the determination of test specifics that are crucial to test quality.

4) Expert advisory panels ought to spend most of their time and energy specifying the populations of knowledge elements that the test items will sample.

5) Any complex achievement can be assessed validly by testing separately the elements of knowledge that make it possible.

6) For measuring educational achievement, simple items are usually more efficient and more discriminating than complex items.

7) While it is proper to include items in the assessment instruments which do not discriminate within sub-groups, it is not wise to disregard more appropriate indices of discrimination in revising and selecting items.

8) To obtain an unbiased picture of how much students are learning of what they are supposed to learn, the test constructors should not select items on the basis of their difficulty indices.

Ebel, R. L. The relation of item discrimination to test reliability. *Journal of Educational Measurement*, 1967, *3*, 125–128.

In this paper, Ebel maintains that in order to achieve high reliability in a test with a given number of items, one must write or select items that are high in discrimination as measured by *D*, the upper-level index of discrimination. Data are presented to support this position.

Ebel, R. L. The value of internal consistency in classroom examinations. *Journal of Educational Measurement*, 1968, *5*, 71–73.

The question discussed in this article is, "Is internal consistency more valuable in a test used for prediction than in one used for assessment?" Ebel notes that Horn believes that it is. He quotes Horn (*Journal of Educational Measurement*, 1966, *3*, 293–5) on this point.

> If a test representatively covers the areas which experts say should be covered (in a bar exam, for example), it makes no difference if the internal consistency reliability is zero, if the test has no variance or if the distribution of scores has a poor form.

Ebel disagrees with this statement of Horn's and asks: If a bar exam yields scores of no variance, what useful purpose does it serve?

Ebel maintains that moderately high internal consistency in assessment devices should be both expected and sought. It should be sought, he claims, because internal consistency tends to insure reliability and, he asserts that in the majority of cases the usefulness of a test score depends on its reliability, regardless of whether the test was intended for assessment or prediction. Ebel gives the following definition of reliability:

Theoretically, reliability is the ratio of true score variance to obtained score variance; operationally, it is the correlation between measurements of the same characteristic obtained from equivalent but independent operations. Reliability of this kind should never be defined as internal consistency, though it may often be estimated by a measure of internal consistency. In principle, a test that is perfectly reliable in the variance-ratio sense or in the correlation of equivalent measurements sense may have zero-internal consistency. But in practice, and in most situations, measures of internal consistency actually do give reasonably good estimates of reliability.

Ebel states that the causes of low internal consistency in classroom tests are often faults in the test items, with the most common faults being ambiguity, indefensibility (of the keyed response), or inappropriateness in difficulty. These are also frequent causes of low reliability. The rejection or revision of such faulty items thus tends to improve both internal consistency and reliability.

The sum of Ebel's argument is that reliability is just as necessary for assessment as for prediction, and that internal consistency estimates of reliability are equally useful in both cases. While he recognizes a difference between the use of test scores for assessment or for prediction, he sees no corresponding characteristic difference between the devices used for these two purposes.

Ebel, R. L. Why is a longer test usually a more reliable test? *Educational and Psychological Measurement*, 1972, *32*, 249–253.

Ebel states that one of the best known properties of the tests commonly used in educational and psychological measurement is that the longer they are, the more reliable are the scores they yield. The explanation for this is given on the basis of two relations:

1) The true component of a score is proportional to the number of equivalent elements that contribute to it.

2) The error component of a score is proportional to the square root of the number of equivalent elements that contribute to it.

The credibility of these two propositions is supported. In view of this, it can be seen that increasing test length increases the true score variance more rapidly than it increases the error variance.

The two propositions are then related to the Spearman-Brown formula. Ebel notes that the differential relations of true scores and errors of measurement to the number of items in a test were pointed out by Gulliksen in 1950.

Emrick, J. A. An evaluation model for mastery testing. *Journal of Educational Measurement*, 1971, *8*, 321-326.

The desirability of criterion-referenced test procedures is noted and an evaluation model based on the following assumptions is presented.

1) The learning of fundamental skills can be considered all or none.

2) Within a single skill test, each item response provides an unbiased estimate of the examinee's mastery status with respect to that skill.

3) Measurement error for a given examinee on a single skill can be of only one type, $\alpha$ or $\beta$.

4) Measurement error occurring on the test can be approximated by calculating the average interitem correlation. The $\phi$ represents the correlation on the item level, and the square of it is the expression for item reliability on a single skill mastery test. $\phi = (1 - \alpha - \beta)/\sqrt{(1 - (\alpha - \beta)^2}$

5) Due to the presence of some measurement error, decision errors will accrue regarding determination of examinee status on the skills being measured. A decision-theoretic approach to this problem suggests that regret due to these evaluation errors can be minimized through a cost-benefit analysis of the variables which comprise the evaluative process.

The resultant mastery criteria algorithm is:

$$K = \frac{\log(\beta/(1 - \alpha)) + 1/n(\log RR)}{\log \alpha\beta/((1 - \alpha)(1 - \beta))}$$

where $K$ = the cut point expressed as a percent score on the test
$\alpha$ = estimated probability of Type I item error
$\beta$ = estimated probability of Type II item error
$RR$ = Ratio of Regret of Type II to Type I decision errors
$n$ = test length (number of items).

Evans, F. R. & Pike, L. W. The effects of instruction for three mathematics item formats. *Journal of Educational Measurement*, 1973, *10*, 257-272.

The authors developed three different instructional programs for three mathematics aptitude item formats to determine the relative susceptibility of each to special instruction. The three item formats were part of the SAT mathematics test. They include:

1) *Quantitative Comparison* (*QC*) item—presents the candidate with two quantities, one in column A and the other in column B. The examinee's task is to compare the magnitude of the two quantities and to mark A if the quantity in column A is larger; B, if the quantity in column B is larger; C, if the quantities are equal; or D, if there isn't enough information to determine the quantitative relationship.

2) *Data Sufficiency* (*DS*) item—presents the candidate with a question followed by two statements, labeled (1) and (2), in which certain data are given. The examinee's task is to decide whether the question can be answered by A, (1) alone; B, (2) alone; C, (1) and (2) together; D, either (1) alone or (2) alone; or E, neither statement alone nor by (1) and (2) together. (The authors consider DS format to be the most complex.)

3) *Regular Mathematics* (*RM*) item—presents the candidate with a problem and five possible solutions. He has to determine which is the correct solution.

The susceptibility of QC items to special instruction was of greatest interest to the authors. The relative susceptibility of geometric & nongeometric items within the three formats was of secondary interest.

Parallel test forms were constructed as pre- and post- measures, and the data were analyzed in a two-way (treatment by sex) multivariate analysis of covariance.

Subjects were male and female high school junior volunteers in 12 schools. In the seven weeks between a pre- and post-test, experimental S's received 21 hours of instruction for one of the three formats; control S's received no special instruction.

Results of the statistical analysis showed that each of the three item formats was susceptible to the special instruction specifically directed toward it. The complex or novel item formats appeared to be more susceptible than the relatively straightforward item format. Female volunteers were found to be slightly less able mathematically at the outset and to benefit somewhat less from the instruction than male volunteers. Mean gains of nearly a full standard deviation obtained by the groups instructed for the complex or novel formats were considered to be of practical consequence and likely to influence admission decisions. The results were consistent for all 12 schools. Although no group received instruction for the SAT-M per se, substantial pre- to post-test gains on that measure were also observed.

Finn, R. H. A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 1970, *30*, 71-76.

The author gives examples of new methods for estimating the reliability of categorical data.

Fleiss, J. L. & Cohen, J. The equivalence of weighted kappa and the interclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 1973, *33*, 613-619.

For appraising reliability, kappa or weighted kappa are useful measures of inter-rater agreement for categorical scales.

Kappa is the proportion of agreement corrected for chance, and scaled to vary from −1 to +1 so that a negative value indicates poorer than chance agreement, zero indicates exactly chance agreement, and a positive value indicates better than chance agreement, with unity indicating perfect agreement. When the investigator can specify the relative seriousness of each kind of disagreement, he may employ weighted Kappa, the proportion of weighted agreement corrected for chance.

This paper establishes the equivalence of weighted kappa with the intraclass correlation coefficient under general conditions (as opposed to the restricted conditions previously established by Cohen).

Weighted Kappa is defined by $K_w = (\bar{D}_c - \bar{D}_o)/\bar{D}_c$ where $\bar{D}_c$ = mean of disagreement expected by chance and $\bar{D}_o$ = mean observed degree of disagreement.

This paper establishes a more general property of weighted kappa. Specifically, that if $v_{ij} = (i - j)^2$ (where $v_{ij}$ denotes the disagreement weight associated with categories $i$ and $j$), and if the categories are scaled so that the first category is scored 1, the second category 2, etc. (which is valid only when the categories may be ordered), then, irrespective of the marginal distributions, weighted kappa is identical with the intraclass correlation coefficient in which the mean difference between the raters is included as a component of variability.

The authors point out that the intraclass correlation coefficient is the special case of weighted Kappa when the categories are equally spaced points along one dimension.

### Frisbie, D. A. Multiple choice versus true-false: A comparison of reliabilities and concurrent validities. *Journal of Educational Measurement*, 1973, *10*, 297-304.

The purpose of this study is to compare the reliabilities and concurrent validities of multiple choice (MC) and true-false (TF) tests that were written to measure understandings and relationships in the same content areas.

Multiple choice items from a widely used battery of achievement tests were changed to true-false format using two different procedures.

1) *Judgmental coversion method* (J)—teachers judged the multiple choice distractor for each item that appeared to be most plausible for making a false statement with the stem.

2) *Discrimination conversion method* (D)—the distractor for each item with the largest lower-upper difference was used to make a false statement with the stem.

Kuder-Richardson formula 20 reliability coeffcients were computed. The reliabilities were then adjusted with the Spearman-Brown formula. Fisher's Z-transformation is not applied to the stepped up reliabilities (Lord discusses this point in *Journal of Educational Measurement*, 1974, *11*, 55-57). Results show that the TF tests were significantly less reliable than the MC tests.

To compare the concurrent validity of the TF tests and the MC tests a Pearson product-moment correlation coefficient was calculated between subtest scores for each of the eight final test forms. the coefficients were corrected for attenuation and the Forsyth & Feldt (1969) statistic was used to generate 90% confidence intervals for the eight disattenuated coefficients. The hypothesis that the disattenuated correlation coefficient does not differ from unity was supported in six of the eight cases.

Frisbie concludes that the TF test did tend to measure the same thing as the corresponding MC test.

Frisbie, D. A. The effect of item format on reliability and validity: A study of multiple choice and true-false achievement tests. *Educational and Psychological Measurement*, 1974, *34*, 885-892.

The purposes of this study are:

1) to determine the concurrent validity of multiple-choice (MC) and empirically-lengthened true-false (TF) tests,

2) to compare the reliabilities of MC and empirically lengthened TF tests,

3) to determine the number of MC and TF items subjects can attempt in a fixed period of time.

A sample of 529 nonurban high school students each responded to one of four test forms which differed in subject matter (natural sciences or social sciences) and item form order (TF or MC). The results showed that the ratio of the number of TF to MC items attempted in the first eight minutes of testing was 3:2. The reliabilities of the MC tests were significantly greater than those of the TF tests. The paper concluded that MC and TF tests designed to measure the same objectives do tend to measure the same characteristic.

Gardner, P. L. Test length and the standard error of measurement. *Journal of Educational Measurement*, 1970, *7*, 271-273.

The author shows that under very general conditions, the standard error of measurement estimated from the Kuder-Richardson formula 20 and Kuder-Richardson formula 21 leads to Lord's observations that the standard error of measurement of a test is directly proportional to the square root of the number of items on the test.

Gilman, D. A. & Ferry, P. Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, 1972, *9*, 205-207.

A 66-item four-response multiple choice test on self-scoring test forms was administered to fifty-four graduate students. Each test was scored by the traditional right-wrong method of scoring tests (RWSM) and also by the self scoring method of counting the number of responses necessary to respond to all items correctly (SSM). This study compared the test reliability of SSM with the reliability of inferred RWSM by using odd-even item correlation coefficients and split-half reliability coefficients (Spearman-Brown Prophesy Formula).

Results indicate that the odd-even correlation coefficient and the split-half reliability coefficient were substantially larger when tests were scored by SSM than when the tests were scored by RWSM.

Glass, G. V. & Wiley, D. E. Formula scoring and test reliability. *Journal of Educational Measurement*, 1964, *1*, 43-47.

A model is proposed for the partitioning of an obtained score for a subject on a multiple-choice test. Deductions from the model were found to correspond to the results of calculations with actual data in a large proportion of cases. Specifically, it was argued that the reliability of uncorrected test scores is generally higher than that for corrected-for-guessing test scores, depending on how well the actual data conform to the proposed model. The greater validity of corrected-for-guessing scores found by other researchers is reflected in the fact that corrected-for-guessing scores lead to a more powerful analysis of variance than uncorrected scores.

Grier, J. B. The number of alternatives for optimum test reliability. *Journal of Educational Measurement*, 1975, *12*, 109-113.

With the total number of alternatives fixed at some constant number, the use of three alternatives per choice was found to maximize three criteria: the power of a test, defined as one minus the probability of getting a perfect score by chance; the discrimination capacity of the test, defined as the number of possible response patterns the test can distinguish between; and the uncertainty index, a measure of the information gained from using the test. The proof used Ebel's (1969) modified version of the Kuder-Richardson formula 21. The expected reliability of a test is maximized, however, only if the number of test items is increased to compensate for the smaller number of alternatives per item.

Hakstian, A. R. & Kansup, W. A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. *Journal of Educational Measurement*, 1975, *12*, 231-239.

A comparison of reliability and validity was made for three testing procedures. A sample of 1028 grade nine students was randomly divided into groups which:

1) responded conventionally to Verbal Ability and Mathematical Reasoning tests,
2) used a confidence-weighting response procedure with the same tests,
3) used the elimination response method.

Data on school achievement criteria were obtained and a similar ability measure was administered to assess criterion-related validity. Elimination test scores showed no increase over conventional scores, in either consistency or stability, and no significant increase in validity. Confidence test scores showed, in some cases, significantly higher reliability than did conventional scores. However, the increase would be matched by conventional tests requiring equal testing time. The confidence scores yielded no increase, and in some cases a decrease in validity. It was concluded that the experimental testing procedures examined are not psychometrically superior to conventional testing.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 93-99.

The author presents a rationale for using classical test construction and analysis procedures when samples of both mastery and nonmastery examinees are employed.

Test and item statistics were computed for three different samples:

a) preinstruction students who represented a nonmastery population,
b) postinstruction examinees who represented the mastery population,
c) a combination of the above two samples.

Homogeneity was estimated using the Kuder-Richardson formula 20, and discrimination indices were computed using $D\%$ and the point biserial correlation for both the postinstruction sample and the combined pre- and postinstruction samples. The author concluded that both logical rationale and empirical evidence support the practice of combining pre- and postinstruction CR test scores for the purpose of examining CR tests and item characteristics.

Hales, L. W. Method of obtaining the index of discrimination for item selection and selected test characteristics: A comparative study. *Educational and Psychological Measurement*, 1972, *32*, 929–937.

The purpose of this study is to determine the relative value of three item validation methods which may be employed by classroom teachers in the selection of items for inclusion in a test. The three methods which were compared are: Flanagan's $r$, Flanagan's $r_c$ computed from proportions which have been corrected for chance success (and having corrected indices of difficulty falling within the range 0.15–0.75, inclusive), and net $D$.

Using the three techniques for item validation, nine tests were constructed (one test by each method for each of three groups). The average overlap between tests of a grade level was 62%.

At each grade level, the $r$ test mean was significantly higher than the $r_c$ test mean. The $D$ test mean fell in between $r$ and $r_c$ test means. For the tests at each grade level, the Kuder-Richardson formula 20 and the odd-even coefficients of correlation for the tests did not differ significantly from each other.

The author concludes from the results of this study that the net $D$ is as good as Flanagan's $r$ and Flanagan's $r_c$ as an index of discrimination to be used in item selection in test construction. Hales notes that since net $D$ may be obtained much more rapidly than either Flanagan's $r$ or $r_c$, the net $D$ should be an appropriate index of discrimination for classroom teachers to use, in conjunction with the index of difficulty, in the selection of items for inclusion on a test.

Hambleton, R. D. & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159–170.

The authors synthesize some of the thinking in the area of criterion-referenced (CR) testing (current in 1973) as well as provide the beginning of an integration theory and method for such testing. They view criterion-referenced testing from a decision-theoretic point of view; thus approaches to reliability and validity estimation consistent with this philosophy are suggested. In order to improve the decision-making accuracy of CR tests, a Bayesian procedure for estimating true mastery scores is proposed. This Bayesian procedure utilizes information about other members of a student's group (collateral information), but the resulting estimation is considered to be criterion-referenced rather than norm-referenced since the student is compared to a standard rather than to other students. The authors contend that in theory, the Bayesian procedure increases the "effective length" of the test by improving the reliability, the validity, and the decision-making accuracy of the criterion-referenced test scores.

Hambleton, R. K., Roberts, D. M. & Traub, R. E. A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement*, 1970, *7*, 75-82.

The purpose of this study is to compare two procedures: differential weighting of response alternatives, and confidence testing—both which have the same goal, assessment of partial knowledge. Comparisons were made in terms of reliability (which was estimated by the split-half technique) and validity (which was estimated by correlating midterm test scores with scores on a final examination.)

1) Differential weighting procedures means that differential scoring weights were assigned to each response alternative to an item rather than a score of +1 for correct answers and 0 for incorrect. In this study weights were chosen to reflect the judged degree of correctness.

2) Confidence Testing procedure refers to any of a variety of procedures which had the examinee indicate his confidence in the correctness of the response alternatives of an item.

The authors use as a baseline for comparison, results obtained by administering a test under conventional directions.

The confidence testing procedure yielded the most valid and the least reliable scores. The second set of differential weights produced scores with the most reliability. Conventional testing procedures produced scores with the least validity and with as much reliability as the scores yielded by the second set of differential weights.

The authors suggest that the results be interpreted cautiously and that they have limited generality.

Hanna, G. S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, 1975, *12*, 175-178.

This study was designed to replicate an investigation about an experimental answer-until-correct (AUC) procedure. Hanna theorizes that an AUC procedure (one in which an examinee continues to respond to each multiple-choice item until feedback signifies that he is successful) would yield scores of greater reliability and validity than would conventional procedures. Results showed that compared to the inferred conventional scores, the experimental scores were more reliable but less valid. Hanna concludes that content validity of achievement tests demands thoughtful deliberation and should not be forsaken in the pursuit of reliability.

Hansen, R. The influence of variables other than knowledge on probabilistic tests. *Journal of Educational Measurement*, 1971, *8*, 9-14.

In probabilistic test and scoring systems, the examinee is required to respond to each of the options of a multiple-choice test with a probability which represents the confidence he has in that option.

The author contends that if the confidence test is functioning properly, the responses made by an individual should be determined principally by what he knows. To the extent that other idiosyncratic traits of the individual influence his responses, the test will be a less valid indicator of knowledge.

The purpose of the study was to seek the relationship between the degree to which examinees display certainty in their responses and certain personality variables. Although proponents of probabilistic testing would expect these correlations to be low, the author found them to be high. He found indications that response style was related to certain aspects of personality.

It was found that individuals do respond to multiple-choice questions with a characteristic certainty that cannot be accounted for on the basis of their knowledge. This certainty is related to scores of both the F Scale and the Kogan & Wallace risk-taking measure.

Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement*, 1971, *8*, 291-296.

The purpose of this study was to determine in what way Guttman weighting affected the internal consistency and the interrelation of the subtests of a multiple-choice objective test. Subtests of the Scholastic Aptitude Test (SAT) were scored first with Guttman weights and then with conventional correction-for-guessing weights. When Guttman weights were used, the internal consistency of the tests increased markedly. The correlation of the two verbal subtests increased to some extent when Guttman weights were used, but the correlation of the two mathematics subtests as well as the intercorrelation of all verbal and mathematics subtests decreased. Differences in the factor structure of the Guttman—and conventionally—weighted subtests were used to explain the result.

The author suggests that further research be done before implementing a Guttman weighting technique on a large scale. Specifically, he suggested that future research show what a Guttman weighted test measures and what effect the Guttman weighting has on validity.

Hively, W., Patterson, H. L. & Page, S. A. A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 1968, *5*, 275-290.

This paper shows that two quite different approaches to achievement testing converge. One is the strong form of educational behaviorism exemplified by B. F. Skinner's work in the area of programmed instruction. The other is the positivistic approach to mental test theory exemplified by Cronbach's work on "Generalizability." Hively *et al* maintain that if behaviorists can analyze non-trivial subject matter into well-defined behavioral classes, generalizability theory promises appropriate and powerful measurement models. Osburn (1968) has named this area "universe-defined-achievement testing."

Data are presented from one of the first applications. The subject matter is mathematics.

A general form, together with a list of generation rules, precisely define the set of all test items which may be taken to represent the diagnostic category. The rules for generating such a set of test items is called an "item form." A collection of item forms constitute a "universe" from which tests may be drawn. A "family" of random-parallel tests (Cronbach, 1963) is defined by a sampling plan over a universe of item forms. A "generalizability study" of the test families was conducted. Three tests were generated from each family. The results show that the relative magnitudes of the components of variance display an extraordinarily consistent pattern across the different test families. It is emphasized that these results were obtained on the basis of purely formal content analysis of the subject matter, without any statistical item selection procedures whatsoever.

In addition to estimates of "relative" stability of test scores expressed by intercorrelation coefficients, a measure of the individual's performance with respect to the universe, which does not require comparison to other individuals for its interpretation, is also obtainable. To get this, we may use the within-person variance to estimate a confidence interval for any individual's "true" or "universe" score, given his observed score on a randomly chosen test. The underlying assumption in all of the above is that the between- and within-person variances are independent of one another.

The authors continue to say that given information about how a person responded to a particular test item, we would expect to be able to predict how he would respond to another, randomly-chosen item from the same item form, but not necessarily how he would respond to an item from a different item form. Predictions from one item form to another should depend on how the items forms are related.

Hively concludes that the data lead one to place only moderate faith in the item forms as categories which represent distinct, homogenous classes of behavior and which thus provide the foundation for detailed diagnosis and remediation. By contrast, it seems paradoxical that the total test scores should have been as reliable as they were.

Horn, J. L. Integration of concepts of reliability and standard error measurement. *Educational and Psychological Measurement*, 1971, *31*, 57-74.

The purpose of this paper is to explicate some of the problems implied by the assumptions underlying derivations of various indices of error of measurement and such coefficients of reliability as the Kuder-Richardson formula 20 and the Kuder-Richardson formula 21 and to indicate some of the practical implications of various proposed solutions.

Horn looks at standard error of measurement and reliability coefficients as they are defined in terms of two random response models. The conclusion is that generally the Kuder-Richardson formula 20 should yield a larger estimate of reliability than the Kuder-Richardson formula 21, and although the difference may be small in many practical situations, the fact of the difference between the two should be kept in mind when considering the standard error of measurement formulae.

In the second section of this paper, Horn looks at some standard error of measurement models. It is noted that different kinds of variability can be represented as "error" in any one of the formulae for reliability or standard error of measurement.

### Horn, J. L. Is it reasonable for assessments to have different psychometric properties than predictors? *Journal of Educational Measurement*, 1968, *5*, 75-77.

Horn contends that the distinction which Nunnally (*Tests and measurements: Assessment and prediction*. New York: McGraw-Hill, 1959) has drawn between assessments and predictors is more than a classificatory-verbal convenience (See Ebel, *Journal of Educational Measurement*, 1968, 5, 71-73.) He believes that it pertains to differences in the psychometric properties of the measurements in question.

His argument is that an assessment device *may* have most all the properties deemed desirable for a predictor device, but that it *need not*. His argument does not question that an assessment should be reliable in the sense that Ebel (*JEM*, 1968, 5) defines this or that internal consistency can be indicative of reliability in this broad sense. The argument is that internal consistency is a secondary consideration for an assessment and may be a counterindication of its adequacy.

If there were *m* areas of, for example, law and each were represented in the bar examination by only one question, we would expect only random correlation among the items. Under these conditions, high internal consistency would be a counter indication of validity.

It is on this basis that Horn argues that evidence on internal consistency is not of primary importance in evaluating an assessment. The primary concern is to ensure that experts will agree that the content of the measurement scale is appropriate for the assessment. A test which meets this requirement might or might not be internally consistent. It is in this sense that evidence on internal consistency can be (which is not to say it usually is) a very secondary consideration.

The essence of this argument is that in some kinds of measurement, validity questions can be approached directly, without much consideration of replication over similar kinds of stimuli and internal consistency in response to these stimuli; whereas in other measurement situations, validity questions—usually the concern is with construct validity—must be approached more indirectly and a concern for internal consistency must come early to our attention.

### Horn, J. L. Some characteristics of classroom examinations. *Journal of Educational Measurement*, 1966, *3*, 293-295.

Horn maintains that the task of constructing a test can be quite different depending on whether the test is conceived of as a predictor or an assessment. In constructing a predictor, one must strive to obtain internal consistency among the items, since this tends to ensure that the test will be reliable and can correlate with a criterion. In constructing an assessment, however, one should strive to obtain representativeness of content, whether or not elements are internally consistent. Here, the principal concern must be with validity in the sense that experts will agree that the items measure what they are supposed to measure. If a test representatively covers the areas which experts say should be covered (in a bar exam, for example), it makes no difference if the internal consistency reliability is zero, if the test has no variance or if the distribution of the scores has a poor form. Yet these must be important considerations in the construction of a predictor.

Huck, S. W. & Bowers, N. D. Item difficulty level and sequence effects in multiple-choice achievement. *Journal of Educational Measurement*, 1972, *9*, 105-111.

It is noted that certain authorities imply that the proportion of examinees who correctly answer a test item is influenced by the difficulty of the immediately preceding item. The theoretical basis for the hypothesis of a "sequence effect" resides in the general area of test anxiety and its effect upon performance. If present, such a "sequence effect" would cause $p$ (an estimate of item difficulty) to misrepresent an item's "true" level of difficulty. This study was undertaken in an attempt to ascertain whether the $p$-level associated with a multiple-choice test item is biased by the difficulty level of the immediately preceding item. A balanced Latin square design was used to rearrange examination items into various forms. The results do not support the "sequence effect" hypothesis. Although the authors recognize that certain limitations preclude the generalization of their findings to all students or to all testing situations, they contend that their results suggest that comments relating to "sequence effects" should be qualified as compared with presently appearing statements.

Ivens, S. H. Nonparametric item evaluation index. *Educational and Psychological Measurement*, 1971, *31*, 843-849.

The purpose of this paper is to develop a nonparametric index for evaluating the effectiveness of dichotomously scored items that takes into account both the difficulty level and the discrimination of the item. The criterion upon which this new index is based is that the best possible item will have a difficulty of .5 and have perfect discrimination.

When $N$ (individuals) is even, $S_i = (8Y_i - 4n_i(N + 1))/N^2$ when $N$ is odd, $S_i = (8Y_i - 4n_i(N + 1))/(N^2 - 1)$ where $Y_i$ is the rank sum of those individuals who passed item $i$ $\Big[ Y_i = X_i' R_N = \sum_{j=1}^{N} j a_j$, where $X_i' = [a_1, a_2 \dots a_N]$ and $a_j = 1$ if the $j^{th}$ individual passed the item and $a_j = 0$ if the $j^{th}$ individual failed the item. $R_N' = [1, 2, 3, \dots, N] \Big]$ and $n_i$ is the number of individuals who passed item $i$.

Ivens shows that the distribution of $S_i$, for each $N$, has a known variance and is symmetrical about zero with maximum and minimum values of one and minus one respectively.

In summary, $S_i$ is an easily computed nonparametric index that:

1) is dependent on item difficulty and discrimination,

2) has a known range and variance,

3) has a significance test for its difference from zero,

4) can be meaningfully compared across different administrations of the same items,

5) can be computed by using either the total score of the test in which the item is contained or an outside criterion.

Jacobs, S. Behavior on objective tests under theoretically adequate, inadequate and unspecified scoring rules. *Journal of Educational Measurement*, 1975, *12*, 19-29.

The effects of two levels of penalty for incorrect responses on two dependent variables:

1) a measure of risk-taking or confidence, using nonsense items
2) the number of response-attempts to legitimate items

were investigated for three treatment groups in a 2x3, repeated measures, multivariate ANOVA design. The treatment groups were composed of *S*'s responding under one of three scoring-administrative rules: conventional Coomb's-type directions and two variants suggested as mathematically more adequate. Results indicate significant differences among groups and across penalty conditions. Implications for criterion-referenced testing are noted as follows.

While most of the effort in the area of criterion-referenced testing has centered around strategies for item and test development, item and test administration, and item and test analysis, the following question remains: What should *S*'s be told when they confront a criterion-referenced test? The results in this study indicate that behavior observed may vary as a function of item difficulty (if one assumes *S*'s responded to nonsense items in a manner similar to the way in which they would respond to very difficult legitimate items), instructions to the *S*, and the penalty for incorrect responses. Even if the effect is constant across all *S*'s (which would permit valid norm-referenced comparisons), there still remains the possibility—in criterion-referenced testing—that we may misjudge whether or not a specified level of performance (the criterion) was achieved.

Jacobs, S. S. Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement*, 1971, *8*, 15-20.

This study was undertaken to determine the effects of two levels of penalty on the unwarranted expression of confidence, the personality correlates of confidence-expression and the effects on test statistics of confidence-weighting. A final examination was administered to seventy-two *S*s under confidence-weighting instructions with two levels of penalty for incorrect responses.

Penalty 1—if the selected option was incorrect, *S*s lost 0, 2, or 3 points depending on the category of confidence selected (guess, fairly confident, very confident).

Penalty 2—here *S*s lost 0, 4, or 6 points, again depending upon the confidence category selected.

A multiple correlation of .39 ($p < .05$) was calculated between four scales of the CPI and the measure of unwarranted confidence.

The measure of confidence-expression used was:

$$\text{CONF} = \left( \frac{\text{number of errors for which maximum confidence was expressed}}{\text{number of errors}} \right) \times 100$$

A two-way anova revealed no significant main effects or interaction effect attributable to level of penalty or sex. Although increased penalty level had no effect on confidence-expression, the test's reliability decreased from .85 to .39, and the correlation between conventional and weighted scores dropped from .88 to .095.

Kansup, W. & Hakstian, A. R. A comparison of several methods of assessing partial knowledge in multiple-choice tests: I. Scoring procedures. *Journal of Educational Measurement*, 1975, *12*, 219-230.

The effects of:

1) logically weighting incorrect item options in conventional tests, and

2) different scoring functions with confidence tests on reliability and validity,

are examined. A group of grade nine students took conventionally-administered Verbal and Mathematical Reasoning tests, scored conventionally and by a procedure assigning degree-of-correctness weights to incorrect alternatives. Some increase in internal consistency, but a slight decrease in stability, resulted from the weighting. Validity for school achievement criteria was unimproved by the weighting and in some cases significantly reduced. Another group took the same test with confidence testing instructions. Five scoring functions—representing four underlying models—were compared. Slight, but nonsignificant, gains in internal consistency and stability were obtained for two functions. No substantial differences in validity were found. It was concluded that:

1) logical weighting with conventional tests is likely to be unprofitable,

2) the simplest scoring function for confidence tests is as effective as more complex ones.

Kleinke, D. J. A linear-prediction approach to developing test norms based on matrix-sampling. *Educational and Psychological Measurement*, 1972, *32*, 75-84.

The author describes a linear prediction approach for estimating total test scores from a sample of items and compares it empirically with the negative hypergeometric distribution method of approximation (Keats & Lord, *Psychometrika*, 1962, *27*, 59-72.).

Knapp, T. R. A note concerning Carver's "A model for using the final examination as a measure of the amount learned in classroom learning." *Journal of Educational Measurement*, 1970, *7*, 51.

The author raises questions about Carver's model for measuring gain.

Koehler, R. A. A comparison of the validities of conventional choice testing and various confidence marking procedures. *Journal of Educational Measurement*, 1971, *8*, 297-303.

The purpose of this study is to investigate various conditions under which the continuous confidence marking (CCM) response method might enhance construct validity. The convergent and discriminant validity of two confidence marking techniques with that of conventional choice testing was compared. The patterns of intercorrelations within the "multitrait-multimethod" matrix provided evidence of construct validity. Achievement in vocabulary, social studies, and science (traits) was measured by a sixty item test containing true-false and five-alternative items (methods). The test was administered to three randomly assigned groups (one for each response system) totaling 535 *S*'s. The results indicate very slight differences in convergent and discriminant validity that favored conventional testing over confidence marking techniques.

No evidence is provided that improvement in construct validity occurs when confidence marking is the response mode. On this basis, the authors suggest the use of the conventional choice test. They do recognize, however, that confidence testing may be valuable for other purposes, e.g., for use as a tutorial device.

Koslowsky, M. & Bailit, H. A measure of reliability using qualitative data. *Educational and Psychological Measurement*, 1975, *35*, 843-846.

The authors note that in many types of research activities, it is necessary to obtain a reliability measure for qualitative or unordered data. The procedures that are presently available cannot handle such data using the classical reliability measures. Finn's (1970) method assumes internal type data, and Goodman and Kruskal's (1954) formula for handling reliability of unordered data is good for only one item at a time. This paper expands the Goodman & Kruskal formula, and discusses an approach for calculating the inter-rater reliability for a series of items across many subjects. The procedure is considered to be analogous to the usual reliability determination for an achievement test or an attitude test.

Krippendorff, K. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 1970, *30*, 61-70.

The author states that the analyst of a recording instrument may wish to obtain the following:

1) An estimate of the reliability of a population of data over all observers in the universe using the recording instrument. This measure is called *data reliability* and can be interpreted as a measure of the confidence in data.

2) An estimate of the extent to which data reliability could be improved if scale values were to be transformed or their definitions were to be modified for the individual observers. This measure assesses the *systematic error* of the recording process, which, together with a measure of the random error may be said to account for the lack of data reliability.

3) An estimate of the reliability associated with each individual observer, often called *individual reliability*. Such an estimate permits the identification of observers who are detrimental to achieving high data reliability. Deviant observers need either more instruction or cannot be employed in the process of collecting data.

4) An estimate of the extent to which each observer is corrigible by further instruction. Such an estimate would assess *systematic observer biases* which together with the individual's *random error* account for lack of individual reliability.

5) Finally, there is needed an indication of the extent to which a random sample of observers agree on the scoring of each unit of recording. This measurement may be called *unit reliability* and allows one to identify sources of unreliability within the set of observations.

Levine, H. G. & McGuire, C. H. The validity and reliability of oral examinations in assessing cognitive skills in medicine. *Journal of Educational Measurement*, 1970, *7*, 62-74.

In order to assess clinical competence, not adequately assessed by written examinations, three types of oral examinations specifically designed to yield information on high level cognitive functioning were employed with 784 M.D. candidates. Orals lasting two and one-half hours were divided into five half-hour examinations:

The first was designed to sample skills in relating effectively to patients and colleagues (role-playing was used).

The second was designed to sample observation and interpretive skills (candidates were presented with series of visual stimuli and were expected to describe pathology and make inferences).

The third, fourth, and fifth examinations tested the candidates problem-solving skills. To insure that examinations were administered and scored properly, detailed instructions were given to examiners and candidates; training sessions were conducted for examiners; to maximize the reliability of the oral examination, instructors utilized a 12-point scale.

Although previous experience showed interrater reliability to be about .60, combined effects of sampling and rating errors reduced reliability across raters and across cases to between .15 & .20.

Since all oral scores were pooled for purposes of certification, pooled scores were used to obtain an estimate of reliability for the entire group of orals.

The Spearman-Brown correction formula yields a reliability estimate of .47 (for four tests) with an average of .18. These results are consistent with the Anova formula developed by Ebel in which each group of candidates who were rated by the same team is considered a block.

They conclude that oral examinations must be used in combination with other data.

The validity of the orals is investigated in the following ways:

1) *Content validity*—by means of systematic observation and questionnaire,

2) *Concurrent validity*—by correlations with two supervisors' ratings of habitual job performance,

3) *Construct validity*—by fact or analysis of test scores and ratings,

4) *Predictive validity*—to be studied in a ten-year follow up. Results of content, construct and concurrent validity indicated that oral tests identified factors not measured by multiple-choice tests and, therefore, significantly improved the relationship between supervisory evaluations and test scores.

Lewy, A. Discrimination among individuals vs. discrimination among groups. *Journal of Educational Measurement*, 1973, *10*, 19-24.

It is noted that achievement tests are used for discrimination both among individuals and among classes. The parameters utilized for selecting items for a test are based on two item statistics: difficulty level and correlation of the item with the total test score.

Lewy contends that although these parameters are useful to maximize discrimination among individuals they do not necessarily discriminate among classes.

He suggests that the intraclass correlation coefficient of a test item can serve as a parameter for selecting items which maximize discrimination among classes. The larger the value of this coefficient, the larger its contribution to the efficiency of the test for discrimination among classes.

Lewy presents an example to illustrate the relation of the intraclass correlation coefficient to the efficiency with which the test discriminates among classes.

He concludes from his example that by employing the intraclass correlation coefficient as a parameter for item selection, one changes the structure of the test and retains in its final version different items than one would have obtained without utilizing this parameter as an item selection criterion.

Lewy, A. & Shavit, S. Types of examinations in history studies. *Journal of Educational Measurement*, 1974, *11*, 35-42.

This study utilizes disagreement among experts in item classification as a source for examining the structural characteristics of the classification scheme.

A sample of 546 questions was collected from the examinations of 11th grade history courses, and a team of history teachers proposed a series of categories by which questions could be classified. The classification scheme included the following eight areas: historical facts, concepts, description of events, analysis of events, processes, cause & effect relationships, connection between areas of life, and evaluation.

An agreement-disagreement (ADAM) matrix (which was not a correlation matrix) was compiled by counting each pair of classification decisions for the 546 items. The matrix reveals that the highest proximity exists between the categories fact and concepts. Guttman's smallest-space analysis (SSA) applied to ADAM suggests that the item-classification scheme reflects two different approaches to testing outcomes of history: cognitive psychology and philosophy of history.

The interpolation procedure yields two distributions with an equal number of raw score points and identical cumulative percentage values in each.

The curve-fitting procedure involves the development of first, second, or higher degree polynomials, using the interpolated or smoothed score points. Standard least square numerical solutions are used to obtain the constants for the polynomials or prediction equations. The accuracy with which a given polynomial reproduces the actual trend of the bivariate distribution is assessed by computing the variance of the errors of estimating $Y$ from $X$.

A Fortran IV program was written to carry out the complete analytical equipercentile equating method.

The authors recognize that their proposed analytical solution is appropriate only in those instances that the graphic one is appropriate. It does not solve the problems associated with a small number of subjects, highly skewed distributions, insufficient data points at the ends of the distributions, etc. The solution should be viewed only as an analogue to the graphic method. It has merit because its results are verifiable, it i fast and inexpensive if done on a computer, it eliminates tedious hand-smoothing, it puts the equipercentile method on an analytic basis similar to the linear method of test equating.

Conclusions:

1) The disagreement pattern in classifying objects according to categories can be utilized to measure the distance between categories.

2) SSA can be used for revealing structural characteristics of models dealing with examination questions.

3) History examination questions can be classified according to categories which differ with respect to level of complexity of cognitive functioning and with respect to the degree of heterogeneity of events dealt with.

4) Differentiation according to heterogeneity of events appears only in connection with categories of high level cognitive functioning.

Lindsay, C. A. & Prichard, M. A. An analytic procedure for the equipercentile method of equating tests. *Journal of Educational Measurement*, 1971, *8*, 203-207.

From an analytical point of view, the graphic procedure for the equipercentile method involves two steps:

1) interpolation or smoothing,

2) extrapolation based on a curve fitting procedure. The authors' proposed method used a linear rule to interpolate the two obtained distributions, then develops functional equations from the interpolated distributions for matching and extrapolation.

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion-referenced testing: Comments on reliability." *Journal of Educational Measurement*, 1972, *9*, 139-140.

Livingston maintains that while the criterion-referenced reliability coefficient may be "unnecessary," as Shavelson *et al* (1972) have claimed, it is useful inasmuch as it gives the user a single number that indicates the reliability of a group of scores in relation to a criterion score. The author states that his CR reliability coefficient provides the answer to the question, "What proportion of the information provided by this test is reliable information?"

Livingston defends the notion that the CR reliability of a group of scores should depend heavily on the difference between the group mean score and the criterion score.

Referring to the hypothetical instructor in Shavelson *et al*'s (1972) example, Livingston states that the fact that the CR reliability formula can be incorrectly applied does not mean that it is useless. Finally, Livingston maintains that the CR reliability coefficient deserves to be called a reliability coefficient because it represents the ratio of "true" to "observed" mean squared deviations from the criterion score, and the CR correlation between alternate forms of the same test and the squared CR correlation between true scores and observed scores. Thus, Livingston contends it is directly related to the repeatability of the measure, when repeatability is assessed in terms of deviations from the criterion score, rather than from the group mean.

Lord, F. M. Formula scoring and number-right scoring. *Journal of Educational Measurement*, 1975, *12*, 7-11.

Lord notes that the assumption that examinees either know the answer to a test item or else guess at random is usually implausible. Number-right scoring directions inform the examinees that to maximize their score it is advantageous for them to answer every item in the test, even if they should have to choose some answers at random. Formula-scoring directions inform the examinees that to maximize their score it is advantageous for them to answer each multiple-choice item whenever they have any valid partial information to guide them in choosing the right answer or in ruling out any of the alternative choices.

Lord makes the following assumption: the difference between an answer sheet obtained under formula-scoring directions, as here defined, and the same answer sheet obtained under number-right scoring directions, as here defined, is only that omitted responses, if any, on the former answer sheet are replaced by random guesses on the latter. This assumption freely permits examinees to use any and all valid partial information available to them. This assumption probably holds best for unspeeded tests.

Under this assumption, formula scoring is found to be clearly superior to number-right scoring. It is noted that the advantage of the formula-scoring over number-right scoring depends on the number of omitted responses. Thus, the advantage will be negligible for high-ability students who know the correct answers and greatest for low-ability students who omit many items. It is suggested that empirical studies to investigate the validity of the new assumption need to be conducted.

Lord, F. M. Quick estimates of the relative efficiency of two tests as a function of ability level. *Journal of Educational Measurement*, 1974, *11*, 247-254.

It is noted that when comparing two tests that measure the same trait or ability, separate comparisons should be made at different levels of the trait or ability rather than an overall comparison. If the two tests are not equally difficult, relative efficiency (R.E.) will differ at different ability levels.

This paper presents a simple approximation formula for R. E. at any specified score level.

The accurate formula to find R. E. of two tests measuring the same trait is:

(formula 1)
$$\text{R.E.}(y,\ x) = \frac{\text{Var}(x|T_x)g_x^2(T_x)}{\text{Var}\ (y|T_y)g_y^2(T_y)} .$$

An approximation to R.E. can be obtained by substituting observed-score distributions $f_x$ and $f_y$ for true-score distributions $g_x$ and $g_y$, and by replacing the ratio of conditional variance by the approximation $n_y x(n_x - x)/n_x y(n_y - y)$.

The approximate formula (2) is:

$$\text{R.E.}(y,\ x) = \frac{n_y}{n_x} \frac{x(n_x - x)}{y(n_y - y)} \frac{f_x^2}{f_y^2}.$$

Using vocabulary sections of seven nationally known reading tests, six practical applications illustrate the adequacy of formula 2 as a convenient approximation to formula 1.

R.E. were computed by formula 2 and by a computer program which gives a good approximation to formula 1.

The conclusion is that the approximation of formula 2 to 1 seems to be adequate.

Lord, F. M. The self-scoring flexilevel test. *Journal of Educational Measurement*, 1971, *8*, 147-151.

It is noted that with "tailored" testing, matching the difficulty of the items with the ability level of the examinee presents great practical complications. In this paper, Lord suggests that these same results can be achieved by modifying the directions, the test booklet, and the answer sheet of an ordinary conventional test. He calls the modified test a *flexilevel test*.

It is assumed that conventional tests are arranged either in the order of difficulty or a rough approximation to this. The general idea of a flexilevel test is simply that the examinee starts with the middle item in the test and proceeds, taking an easier item each time he gets an item wrong, a harder item each time he gets an item right. He stops when he has answered half the items in the test. The answer sheet must inform the examinee whether each answer is right or wrong.

The high-ability examinee who does well on the first items he answers will automatically be admistered a harder set of items than the low-ability examinee who does poorly on the first items. Within limits, the flexilevel test automatically adjusts the difficulty of the items administered to the ability level of the examinee.

This result is not achieved without some complication of the test administration. However, the complications are minor compared with those arising in other forms of tailored testing.

Lord, F. M. Variance stabilizing trasformation of the stepped-up reliability coefficient. *Journal of Educational Measurement*, 1974, *11*, 55-57.

Lord states that since the stepped-up reliability coefficient does not have the same standard error as an ordinary correlation coefficient, Fisher's $z$-transformation should not be applied to it. He suggests appropriate procedures.

Macready, G. B. & Merwin, J. C. Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement*, 1973, *33*, 351-360.

This paper considers the nature of the relationships among items within item forms and how these relationships compare with an ideal case for diagnostic tests in which if a person gets one item within an item form right, then he should get all items within the item form correct.

The results suggest that, in most cases, item forms which generate items of moderate difficulty can be used to obtain relatively homogeneous sets of items of equivalent difficulty for a defined population of subjects. Such item forms provide sets of items superior to those which would be expected if item difficulties alone were used to group items into sets. This suggests that the means used in defining the replacement-set structures by attempting to objectify the intuitive categories ordinarily used by teachers in constructing diagnostic tests is at least a reasonable first effort.

The results also suggest a basis for identification of item forms which will generate homogeneous items of similar difficulty. Using this information, it is possible to determine whether the breadth of an item form is appropriate and if not, identify changes which will lead to an item form of more useful breadth.

Marso, R. Test item arrangement, testing time, and performance. *Journal of Educational Measurement*, 1970, *7*, 113–118.

Two experiments were conducted to determine if a relationship exists between test item arrangements and student performance on power tests.

The following hypotheses were tested:

1) item arrangement based upon item difficulty will not influence student performance or required testing time when tests are administered as power tests,

2) item arrangements based on similarity of content covered or on order of class presentation will not influence student performance or required testing time when tests are administered as power tests,

3) the item arrangement factor will not interact with the test anxiety factor.

Results indicate that arranging items according to difficulty has little or no effect upon either required testing time or upon student performance on power or achievement tests. Students with greater or less measured anxiety did not perform differently when item difficulty arrangements varied, but those with greater anxiety performed less well on the examinations.

From experiment two, results indicate that the item presentation formats did not influence student performance on the final examination nor did this factor interact with the levels of test anxiety. Again, students with high levels of test anxiety performed less well on classroom examinations.

Masters, J. R. The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, 1974, *11*, 49-53.

The purpose of this study is to investigate the relationship between the number of response categories employed and internal-consistency reliability of Likert-type attitude questionnaires. To date, contradictory evidence has resulted in no clear conclusion about the relationship between the number of response categories and reliability.

A 17-item Attitude toward Educational Traditionalism questionnaire and a 22-item Attitude toward Educational Progressivism questionnaire were each scaled with 2, 3, 4, 5, 6 & 7 categories. They were administered to graduate students in education and coefficient alpha reliabilities were obtained for each sample for each of the six scalings and for each total questionnaire.

Progressive questionnaire reliabilities increased substantially as a function of increasing the number of categories, but reliability of the Traditional questionnaire proved independent of the number of categories employed.

The obtained reliabilities for larger numbers of categories on the Progressive quesionnaire were found to be much higher than would be predicted through the use of the Spearman-Brown formula utilizing data for two categories.

The most reasonable explanation for the different results for the two questionnaires was found in examining the two-category total score distribution of each questionnaire. As the number of categories increased, the variability of the Progressive questionnaire increased greatly. Results indicate that in situations where low total score variability is achieved with a small number of categories, reliability can be increased through increasing the number of categories employed. In situations where opinion is widely divided toward the content being measured, reliability appeared to be independent of the number of response categories.

Maxwell, A. E. The effect of correlated errors on estimates of reliability coefficients. *Educational and Psychological Measurement*, 1968, *28*, 803-811.

The procedure whereby the "reliability" coefficient of a test can be derived by analysis of variance is reviewed. The assumptions underlying the analysis of variance model are noted and it is shown that if the error terms in the model are not independent then the estimate of the reliability coefficient will be biased, and in most commonly occurring cases will be an overestimate.

Menne, J. W. & Tolsma, R. J. A discrimination index for items in instruments using group responses. *Journal of Educational Measurement*, 1971, *8*, 5-7.

Item discrimination for instruments used to measure characteristics by means of group responses is stressed. It is argued that a percentage of the total sum of squares which is due to groups (between groups) can appropriately be used as an index of item discrimination. The measurement units of concern are the consensual responses made to the items by members of the group or groups in question.

The items selected must be capable of:

1) eliciting similar responses from members of the same group

2) eliciting different responses from members belonging to a different group when the groups in question have been exposed to or have perceived dissimilar conditions.

For discrimination, the within-group variance should be low in relation to the between-group variance.

It is noted that the *F* statistic or ratio of between to within MS is not an entirely suitable index of item discrimination because it is influenced by sample size. The percentage of Total *SS* due to between groups is a suitable index since it is independent of sample size.

The authors suggest that the efficacy of measuring instruments which use group responses can be improved in two ways:

1) inform the user of the minimum situation (i.e. the number and size of the groups for which the instrument was developed),

2) adopt an item selection criterion which will allow the instrument to be used effectively in the minimum practical situation for which its use was intended.

Some may think that item discrimination for every item is not too important in the group measuring situation, since the scale scores will generally discriminate. However, the measurement of group responses is not a close parallel to the measurement of individuals. In the measurement of characteristics by group responses:

1) item scores are generally regarded as important,

2) scale scores are generally based on as few as 5-15 items.

Muller, D., Calhoun, E. & Orling, R. Test reliability as a function of answer sheet mode. *Journal of Educational Measurement*, 1972, *9*, 321-324.

The authors' investigation is aimed at evaluating the effects of answer sheet mode on reliability of measurement at the middle and upper elementary grade levels.

It was found that response mode significantly affected both variability and magnitude of error rate at each of the three grade levels. Because the test was constructed in such a way that answers to almost all the test items were common knowledge to the majority of students, errors on the test almost always reflected errors in marking rather than knowledge. Thus, the authors contended, error rate can be taken as an index of reliability. The results indicate that the use of the separate answer sheets does result in decreased test reliability.

Since not only level of performance is influenced by answer sheet mode, but also reliability, it is recommended that test standardization information specify the answer sheet mode that was used to evaluate reliability and to establish norms.

Niedermeyer, F. C. & Sullivan, H. J. Differential effects of individual and group testing strategies in an objectives-based instructional program. *Journal of Educational Measurement*, 1972, *9*, 199-204.

Prior to this study, it was determined that the typical, three-choice, selected-response tests for a first-grade reading program were inappropriate for assessment purposes since many children scored well on these unit tests during the year, but did not reach criterion on an end-of-year, constructed-response posttest. To resolve this problem two other types of tests were developed and compared to the three-choice test.

Ten first-grade teachers in an objectives-based reading program utilized on a biweekly basis three types of criterion tests:

1) individually administered, constructed-response tests (the teacher administered these by asking a child to come up and read the word on the card),

2) group-administered, selected-response tests with three choices per item,

3) group-administered, selected-response tests with four choices per item. This test was developed directly from the existing three-choice test by simply adding a fourth distractor to each item. The fourth distractor was generated in such a way as to make the item more difficult.

Scores on these three tests and scores on an end-of-year, constructed-response posttest were collected on a sample of 40 *S*'s for each type of test. The results indicated that the three-choice, selected-response test often utilized in programs of this type does not provide an accurate indication of end-of-year achievement for many children. The authors do not recomment its continued use.

Both the individually-administered, constructed-response tests and the four-choice, selected-response tests provided scores that accurately predicted end-of-year performance. They both produced scores that were lower than for the three-choice test and thus allowed teachers more easily to identify pupils who required remediation.

Nitko, A. J. & Feldt, L. S. A note on the effect of item difficulty distributions on the sampling distribution of KR-20. *American Educational Research Journal*, 1969, *6*, 433-437.

To investigate the possibility of the effect of the item difficulty distribution on the sampling distribution of KR-20, two extreme types of distributions of $\phi_j$'s were constructed:

1) a uniform distribution over the range $\phi_j$ = .20(.05).80

2) a highly concentrated distribution in the neighborhood of $\phi_j$ = .50 with the range $\phi_j$ = .45(.05).60.

Several Monte Carlo experiments were conducted to examine the effect of these item difficulty distributions on the sampling distribution of KR-20. Ten distributions of KR-20 were obtained—two under each of five levels of population reliability.

Data suggests that the effect of these two extremes of item difficulty distributions is minimal. Some of the small differences in percentiles which exist is attributed to sampling error and to the fact that the population reliabilities were not precisely equal for the two tests.

The authors contend that the data provide strong evidence that the form of the distribution of item difficulties has little effect on the sampling distribution of KR-20.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, *28*, 95-104.

This paper concerns the explicit definition of the universe of content, and the stratified random sampling of items from the universe of content so defined.

A universe defined test is a test constructed and administered in such a way that an examinee's score on the test provides an unbiased estimate of his score on some explicitly defined universe of item content. Two requirements for test construction are:

1) all items that could possibly appear in the test should be specified in advance,

2) the items in a particular test should be selected by random sampling or stratified random sampling from the universe of content.

Osburn's approach to defining a universe of content is to analyze the content area into a hierarchical arrangement of item forms and to develop a program for a digital computer that would compose item sentences given a suitable vocabulary and structural codes for the item forms.

An item form has the following characteristics:

1) it generates items with a fixed syntactical structure,

2) it contains one or more variable elements,

3) it defines a class of item sentences by specifying the replacement sets for the variable elements.

The principal advantage of item forms analysis is that it seems possible to characterize the universe of content as an abstract system while maintaining an unambiguous link between the system and the actual items that appear on any form of the test.

Osburn's treatment of item forms analysis draws on the basic features of Hively's approach with more emphasis on the hierarchical arrangement of item forms into a generalized system. The method is the reverse of Gagné's task analysis in that it proceeds from the general to the specific with an emphasis on the abstract system rather than on specific task elements.

The author contends that the abstract system together with the item generating program satisfies the properties of a universe defined test.

It is the author's stance that the mental test model that has been so successful in aptitude testing is not appropriate for across the board application to achievement testing.

Theoretical implications of a universe defined test included:

1) *Reliability Theory:*

In a universe defined test, a particular test or item sample becomes relatively unimportant and interest is focused on the universe of content. Concer is not focused on a specific test but rather with a procedure for estimating an individual's true score on a universe of content.

Classical mental test theory involving assumptions of test equivalence has been divorced from test content and true score content has been little more than a statistical fiction. The theory of generalizability (Cronbach *et al*) which assumes random or stratified random sampling of test conditions as a starting point has linked reliability theory with test content. Universe defined tests can be made to satisfy rigorously the assumptions of generalizability theory and constitute a practical means for implementation of the theory. In generalizability theory, the concept of the universe true score becomes meaningful.

2) *Validity Theory:*

For a universe defined test, what the test is measuring is operationally defined by the universe of content as embodied in the item generating rules. It should be possible to keep separate the concept of what a test is measuring from the concept of the extent to which the responses of a person sample to the universe of content are related to their responses to other classes of stimuli (construct validity, etc.). In response to Ebel's (*American Psychologist*, 1961, *16*, 640-647.) plea for the operational definition of measurement procedures, Osburn suggests that the most important requirement for the operational definition of a test is the specification of the universe of content.

3) *Item Analysis:*

The author suggests that item analysis techniques be redefined if we are to preserve the idea of random sampling from a specified universe of content. Any decision to exclude items based upon item analysis data must result in a redefinition of the universe of content.

4) *Normative Data:*

The percent correct score of a universe defined test is meaningful (Ebel points out that to be meaningful, any test score must be related to test content as well as to scores of other examinees) because it is related to test content. On most psychological tests, a percent correct score is meaningless because the universe of content is not completely specified and random sampling is neglected.

5) *Matched vs. Unmatched Data:*

Matched—the same sample of items is administered to each subject in the person sample.

Unmatched—the items are randomly sampled for each subject. If the investigator is interested in the absolute score of an individual, it does not matter whether or not data are matched or unmatched. If he wants relative scores for individuals, matched data are required. If he wishes to estimate the mean for a group of persons, and proposes to generalize over persons and items, unmatched are preferred.

Osburn, H. G. The effect of item stratification on errors of measurement. *Educational and Psychological Measurement*, 1969, *29*, 295-301.

This paper shows that, in the case of matched item tests, the reduction in errors of measurement for tests constructed by stratified sampling as compared with tests constructed by random sampling from an indefinite population of items, is a simple function of the variance of the difference between pairs of strata true scores. For unmatched item tests, the reduction in errors of measurement due to stratification is a function of the variance (across strata) of the strata mean true scores plus the variance of the difference between pairs of strata true scores.

These results predict that, in the case of matched item tests the largest reductions in errors of measurement will result from stratification on item content rather than item difficulty while for unmatched item test just the opposite is true.

Owens, R. E., Hanna, G. S. & Coppedge, F. L. Comparison of multiple-choice tests using different types of distractor selection techniques. *Journal of Educational Measurement*, 1970, *7*, 87-90.

The purpose of this investigation is to compare the effectiveness of three item-construction techniques in producing a multiple-choice geometry test of maximal concurrent validity. The three procedures used to develop multiple-choice item distractors are:

1) Judgmental method—the item authors supplied distractors that they believed would be plausible.

2) Frequency method—the test was first administered in completion form and then the most frequent examinee errors were used as distractors when the items were cast in multiple-choice format.

3) Discrimination method—the test was first administered in completion form and then the examinee errors that best discriminated between the high and low scoring students were selected for multiple-choice distractors.

Scores on each of these three multiple-choice tests were correlated with scores on a 17-item, 20-minute geometry completion test of parallel numeric and algebraic content. Matched triads, with 558 subjects in each group, were used. No significant differences in validity were found among the tests.

Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement* 1969, *6*, 1-9.

During the past several years measurement and instructional specialists have distinguished between *norm-referenced* and *criterion-referenced* approaches to measurement. More traditional, a norm-referenced measure is used to identify an individual's performance in relation to the performance of others on the same measure. A criterion-referenced test is used to identify an individual's status with respect to an established standard of performance. This discussion examines the implications of these two approaches to measurement, particularly criterion-referenced measurement, with respect to variability, item construction, reliability, validity, item analysis, reporting, and interpretation.

Pyrczak, F. Validity of the discrimination index as a measure of item quality. *Journal of Educational Measurement*, 1973, *10*, 227-231.

The purpose of this study is to investigate the validity of the item discrimination index. Two parallel forms of an arithmetic-reasoning test were constructed. The items were designed to vary with respect to nine item-writing characteristics. On the basis of responses of 364 examinees, a discrimination index was computed for each item. Three judges independently rated the items using a checklist of the nine characteristics. The average of the judges' ratings for each item was used as the criterion for determining the validity of the indices. The findings indicate that the discrimination index appears to be a valid measure of the quality of the multiple-choice items employed in this study.

Raffeld, P. The effects of Guttman weights on the reliability and predictive validity of objective tests when omissions are not differentially weighted. *Journal of Educational Measurement*, 1975, *12*, 179-185.

The effects of constant weights for omissions using the Guttman system was investigated in this study. The results suggest that a constant weight, equal to the mean of the *K* item alternative weights, will produce increases in internal consistency. These increases are considerably less than those found with differentially weighted omissions. A slight increase in predictive validity was found for Guttman weighted scores using a constant omission weight. Consistent decreases in predictive validity resulted when differential weights for omissions were used.

The author concludes that this study supports the contention that a Guttman-weighted objective test can have psychometric properties that are superior to those of its unweighted counterpart, as long as omissions do not exist or are assigned a value equal to the mean of the *K* alternative weights.

Ramsay, J. O. True score theory: A paradox. *Educational and Psychological Measurement*, 1971, *31*, 715-719.

In classical mental test theory, if there is no a priori reason for accepting the statement, "there is no platonic true score," then it is usually not unreasonable to define true score as the expected value of observed score. Ramsay attempts to show that there are consequences of this assumption.

Reliability is defined as $\varrho = \sigma_t^2/(\sigma_t^2 + \sigma_e^2)$ where $\sigma_e^2 = \sigma_x^2 - \sigma_t^2$. By using a fundamental theorem about variance and noting that $E(x|t) = t$, the variance of observed score for a particular true score is limited. In order to see how different from zero this lower limit on reliability may be in practice, Ramsay proposes the beta distribution be used as a model for the distribution of true score.

The result is a "realistic" lower bound on reliability as a function of true score mean and variance. An example is given which expresses the lower bound on reliability as a function of true score standard deviation.

The author notes three ways out of this paradox.

1) Work only with scores transformed so as to be distributed on an infinite interval. This, he points out, seems to make the concept of true score even more artificial than when it was defined to be the expected observed score.

2) Replace this assumption with some new ones. The danger here is that the resulting test score theory will be stronger and contain even more parameters than can be handled computationally and theoretically.

3) Abandon the whole enterprise of describing test score behavior out of a predictive context and rely on standard statistical methodology to relate one test to another. The author notes that this is a radical approach which few may favor.

Ramseyer, G. C. & Cashen, V. M. The effect of practice sessions on the use of separate answer sheets by first and second graders. *Journal of Educational Measurement*, 1971, *8*, 177-181.

The purpose of this study was to determine the effect of formal practice sessions on the ability of first and second graders to use separate answer sheets on the California Test of Mental Maturity. Academically, the 79 *S*'s were above average. The CTMM was administered twice to all subjects, once employing the test booklet for answers and once employing a separate answer sheet preceded by formal practice session. Significant mean raw score differences between the two formats of 10.30 and 7.19 were obtained for *S*'s in grades one and two respectively in favor of the booklet format.

Results indicate that above average pupils in grades one and two are unable to utilize a separate answer sheet effectively even with prior practice sessions in the use of this format.

It may well be that extended formal practice sessions would give first and second graders the necessary skills to enable them to use separate answer sheets effectively. However, the extra expenditure of time may not be worth the effort.

Reid, J. C. Printed comments with item analysis. *Journal of Educational Measurement*, 1970, *7*, 159-160.

Reid suggests that the teacher who lacks training in measurement often ignores such item analysis options like discrimination and difficulty indices. However, item statistics do help the measurement expert who has developed a mental decision matrix that guides his revision of the item.

Reid also suggests that a way of assisting the teacher to improve the test items is to incorporate the expert's mental decision matrix into a computer program designed to compute item statistics and let the computer print not only the statistics, but also a running commentary in English alongside the item statistics.

These comments should provide directions to the classroom teacher for the improvement of the items. They should be flexible to reflect the differing decisions required by norm-referenced items and criterion-referenced items.

If the printed remarks comment on the relative effectiveness of the instructional message as well as the item itself, then written comments may fit well into an improvement of the instructional system and will help the unit to implement the improvement of instruction.

Reilly, R. R. & Jackson, R. Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement*, 1973, *10*, 185-194.

Item options of shortened forms of the GRE Verbal and Quantitative tests were empirically weighted by two variants of a method originally attributed to Guttman. The first variant method was internal consistency keying and the second was parallel forms keying. The purpose of this study is to provide evidence of the effects of empirical option weighting on the reliability, internal consistency, validity, and factor structure of a standardized academic aptitude test. When compared with formula scores, it was found that tests scored with the empirical weights were more reliable but less valid when correlated with undergraduate GPA. A factor analysis revealed large increases in variance accounted for by the first factor. It is suggested that the weighting procedures used tended to capitalize on omitting behavior which, although a highly reliable tendency, may be invalid.

Rippey, R. M. A comparison of five different scoring functions for confidence tests. *Journal of Educational Measurement*, 1970, *7*, 165-170.

Scores for confidence tests were computed using the following scoring functions:

1) probability assigned to the correct answer,
2) logarithmic function,
3) spherical function (both 2 & 3 maximize a student's score if and only if he doesn't guess),
4) Euclidian function,
5) inferred choice (analogous to conventional multiple-choice scoring: one point if maximum confidence is assigned to the correct option, otherwise nothing.)

The reliabilities of these five scoring functions were estimated using Hoyt's procedure. The simplest function, (1) probability assigned to the correct response, had the highest reliability most of the time. Function (1) led to the most reliable scores and function (5), conventional choice scoring, led to the least reliable scores. The Euclidean function, which is similar to function (1) in its scoring results, produced a comparably high reliability. The spherical, logarithmic, and inferred choice functions did not fare so well.

The data suggest that in the absence of information about the scoring system, subjects assign their confidence in multiple-choice responses on the basis of the intuitively simplest payoff model, and that reliability decreases as scoring functions generate item scores which are progressively discrepant from scores generated by the simplest model.

Rosner, J. Language arts and arithmetic achievement, and specifically related perceptual skills. *American Educational Research Journal*, 1973, *10*, 59-68.

The purpose of this paper is to argue that primary grade reading and arithmetic competencies are closely related to specific and different perceptual skills, namely reading to auditory perception and arithmetic to visual perception.

Three sets of scores were compared: Stanford achievement test, visual perceptual test (VAT) and an auditory perceptual test (AAT). Partial correlations were calculated for AAT and achievement, controlling on VAT; and for VAT and achievement, controlling on AAT.

Results indicate that AAT scores account for significantly more of the variance in language arts subtest scores than do VAT; the reverse was true in accounting for the variance in arithmetic scores.

The data indicate that for children in this study, there were significant relationships between reading achievement and auditory perception, and between arithmetic achievement and visual perception.

It is suggested that instructional programs for primary grade children be based upon the strengths and deficits of their perceptual skills.

Shavelson, R. J., Block, J. H. & Ravitch, M. M. Criterion-referenced testing: Comments on reliability. *Journal of Educational Measurement*, 1972, *9*, 133-137.

In this comment, Livingston's reliability coefficient for criterion-referenced measures is examined and some considerations for determining the reliability of criterion-referenced measures is discussed. The authors find Livingston's statistic to be different from conventional reliability statistics and suggest it be called something other than "reliability." For criterion-referenced measures which satisfy the assumptions underlying the classical test theory model, the authors contend that conventional reliability statistics are appropriate. For measures with underlying multi-dimensional traits, classical test theory may be used for estimating the reliability of homogeneous subscales. When the confidence interval about a student's score includes the criterion, the authors suggest obtaining additional evidence about that student before determining whether or not his performance exceeds or fails to exceed the specified criterion. One method for acquiring additional evidence is two-stage sequential testing.

Shavelson, R. J. & Stanton, G. C. Construct validation: Methodology and application to three measures of cognitive structure. *Journal of Educational Measurement*, 1975, *12*, 67-85.

The authors reviewed certain construct validation methodology and applied some of it to the problem of validating construct interpretations of measures of cognitive structure. Their review covered:

1) construct definitions and their implications for measurement operations and interpretations,

2) three kinds of methods for examining construct interpretations of test scores: logical analyses, correlational techniques, and experimental techniques.

Cognitive structure was defined and implications of the construct definition for measures of cognitive structure were examined. The results of two studies which dealt with the convergence of measures of cognitive structure were reported.

Shoemaker, D. M. Note on the attenuating effect of zero-variance items on KR-20. *Journal of Educational Measurement*, 1969, *6*, 255-256.

Kuder-Richardson reliability formulas 20 and 21 should not be applied to tests containing items either answered correctly or incorrectly by all examinees. The extent to which KR-20 is attenuated by zero-variance items is derived.

**Shoemaker, D. M. Standard errors of estimate in item-examinee sampling as a function of test reliability, variation in item difficulty indices and degree of skewness in the normative distribution. *Educational and Psychological Measurement*, 1972, *32*, 705–714.**

Some procedural guidelines are available to aid the researcher in determining the most appropriate number of subtests, number of items per subtest, and number of examinees per subtest. Conspicuous by its absence in a series of investigations was a systematic examination of the effect on standard errors of estimate due to variations in test reliability. The investigation described in this article was primarily designed to remedy this situation. Additional parameters considered were the variance of item difficulty indices $\sigma_p^2$ and degree of skewness in the normative distribution. The parameters estimated were the mean test score $\mu$ and the standard deviation of test scores $\sigma$.

**Shoemaker, D. M. & Osburn, H. G. A simulation model for achievement testing. *Educational and Psychological Measurement*, 1970, *30*, 267–272.**

A model was developed by the authors that simulates the administration of a single test item to a single examinee. The result is a simulation model of great flexibility for the sampling of both items and individuals.

To obtain type-12 sampling (Lord, F. M. *Psychometrika*, 1955, *20*, 193-200) (random items, subjects, & occasions) three features were simulated by the model.

1) The test items: a set of $k$ items must be selected from an item population. Each item must have a difficulty level and a content reference.

2) The examinee: a person with a specified ability level must be randomly selected from a population of people in which the ability under consideration is normally distributed.

3) Testing of examinees over items: does the individual pass or fail each item in the test?

The model assumes a normally distributed standardized latent ability continuum. The probability of an examinee answering an item correctly is a normal-ogive function of his ability level.

The authors employed their simulation model to empirically study certain estimators, the gamma and gamma-stratified coefficients, of test reliability.

**Simon, G. B. Comments on "Implications of criterion-referenced measurement." *Journal of Educational Measurement*, 1969, *6*, 259–260.**

The purpose of this article is to comment on a paper by Popham & Husek (*Journal of Educational Measurement*, 1969, *6*, 1-9). According to Simon, the distinction between criterion-reference and norm-reference applies not to the nature of the test or to the content or form of the items, but concerns the interpretation and use of the scores from the test. Thus, to keep a clear distinction when referring to criterion-reference or norm-reference, the term *tests* should be dropped and replaced by the term *scores* or *measures*.

Simon postulates that Popham and Husek suggest the use of a confidence interval around the individual score because test-retest correlations may be low because of limited variability. Simon knows of no systematic way of deriving a confidence interval that is not based on variability.

In discussing item analysis, Popham and Husek suggest modifying the use of discrimination indices. Simon asserts that this should apply to any achievement test or any test where the fundamental validity is content validity. He maintains that item discrimination statistics should not determine the content of the test—even if norn-referenced scores are to be used.

In Popham and Husek's discussion of negatively discriminating items, they state that although it may be that some deficiencies in the instruction cause the result, it is more likely that the item is deficient. Simon notes that such a situation may not necessarily represent an item deficiency. Such items may occur when the relationship between the amount of knowledge and performance on the item is nonlinear. Simon concludes that the use of criterion-referenced scores is appropriate to programmed instruction, specified behavioral objectives, formative evaluation and whenever mastery of subject matter or of skills is of prime concern.

## Sirotnik, K. An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement*, 1970, *30*, 891-908.

Following a brief discussion of the methodology of matrix sampling, this paper attempts to demonstrate the following points:

1) Matrix sampling can be viewed as a simple two factor, random model analysis of variance design, the matrix sampling formulas for estimating the mean and variance being simply the point estimate formulas for estimating components of the underlying linear model.

2) These formulas can be based on the weakest possible set of assumptions, viz., random and independent sampling of examinees and items. No assumptions about the statistical nature of the data need be made.

3) The literature is unclear with respect to the effect of the above sampling assumptions on multiple matrix sampling in the estimation of the mean and especially the variance.

4) Of the three alternative procedures suggested to deal with negative variance estimates in multiple matrix sampling—equating the negative estimates to zero, Winsorizing the distribution of estimates, or treating all estimates alike regardless of sign—the third procedure appears to be the most promising. A simulation study is necessary to determine the shape of the small sampling distribution of variance components for matrix sampling as well as the relative efficiency of the three methods for handling negative estimates.

## Sirotnik, K. Estimates of coefficient alpha for finite populations of items. *Educational and Psychological Measurement*, 1972a, *32*, 129-136.

This paper attempts to investigate implications for finite and known item populations of classical test theory and the alpha coefficient among items in paper-and-pendil testing. Finite sampling formulas for $\alpha$ are derived and conceptual problems relating to the treatment of the examinee-item populations are discussed.

The following was shown:

1) An exact estimate of $\alpha$ is possible only in the infinite case; the estimate of $\alpha$ in the infinite case is bounded below and above.

2) If error of measurement variance is conceptualized only as examinee-item response variability, it can not be exactly estimated in either the finite or infinite case. It can be overestimated by $MS_{EI}$.

3) If error of measurement variance is conceptualized as a residual variance obtained by pooling error and interaction components, it can be exactly estimated only in the infinite case by $MS_{EI}$. The exact estimate of error of measurement variance conceptualized in this way in the finite case is bounded below and above by $(1 - (m/M))MS_{EI}$ and $MS_{EI}$ respectively.

Sirotnik, K. On "Estimates of coefficient alpha for finite populations of items." *Educational and Psychological Measurement*, 1972b, *32*, 1025.

Conceptual errors in the earlier article are pointed out by the author.

Sirotnik, K. & Wellington, R. Scrambling content in achievement testing: An application of multiple matrix sampling in experimental design. *Journal of Educational Measurement*, 1974, *11*, 179-188.

This study is designed to research the question of scrambling item content in the construction of achievement tests, in order that general implications could be drawn for both examinee and item populations. To achieve this generality, the methodology of multiple matrix sampling was combined with a simple two-group experimental design: a random group of eighth graders responded to mathematics, science, social studies, reading and language arts achievement items organized in a scramble (random) test format, while another random group responded to the same items organized in a fixed (segregated by subject matter) test format. The results indicate that scrambling cognitive test items has minimal or no effect on mean examinee test performance or on any of the other parameters included in the analysis.

Slakter, M. J., Koehler, R. A. & Hampton, S. H. Grade level, sex, and selected aspects of test-wiseness. *Journal of Educational Measurement*, 1970, *7*, 119-122.

Test-wiseness (TW) has been defined as "a subject's capacity to utilize the characteristics and formats of the test and/or test taking situation to receive a high score."

The purpose of this study is:

1) to construct TW measures suitable for use in grades 5-11,

2) to administer the measures to students in grades 5-11 in order to observe the relation of TW with grade level and/or sex.

The following TW behaviors were chosen: The examinee should be able to:

1) select the option which resembles an aspect of the stem,

2) eliminate options which are known to be incorrect and to choose among the remaining options,

3) eliminate similar options, i.e., options which imply the correctness of each other,

4) eliminate those options which include specific determiners.

A sex by grade multivariate analysis was performed on the four subscale scores. Grade effects were significant at the .05 level. Neither the sex effects nor sex by grade interaction effects were significant at the .05 level. There was an increase in TW over grade level.

Smith, R. B. An empirical investigation of complexity and process in multiple-choice items. *Journal of Educational Measurement*, 1970, 7, 33-41.

In this study, sets of test items were constructed in which there was an attempt to hold content as constant as possible and systematically vary process according to Bloom's *Taxonomy* rationale, (i.e., an "application" test, an "analysis test, etc.) Hierarchical Syndrome Analysis was used to examine the question of whether the various item types can or should be combined. (Hierarchical Syndrome Analysis is a method of classifying people, institutions, or other elements based upon the statistical distance between them.) Each test contained items related to eight physical science principles. The data indicate possible ways of combining *Taxonomy* item types based on the psychological distance between the categories.

Solomon, A. The effect of answer sheet format on test performance by culturally disadvantaged fourth grade elementary school pupils. *Journal of Educational Measurement*, 1971, *8*, 289-290.

This study was initiated to determine whether or not culturally deprived youngsters are penalized by the type of answer sheet used in examinations. One hundred-sixteen fourth graders enrolled in an inner city school receiving Title I funds were randomly assigned to one of three answer sheet formats for the Reading section of the Metropolitan Achievement Test: response in test booklet, response on separate non-machine scorable form, and response on separate machine scorable form.

Answer sheet format was shown to have no effect on the test performance of culturally deprived fourth grade elementary school students.

These results are in concert with those of Gaffney and Maguire (*Journal of Educational Measurement*, 1971, *8*, 42-44) and Cashen and Ramseyer (*Journal of Educational Measurement*, 1969, *6*, 155-158) and do not agree with Hayward's (*Educational and Psychological Measurement*, 1967, *27*, 997-1004).

Stafford, R. E. The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement*, 1971, *8*, 275-277.

Based on Gulliksen's (*Theory of Mental Tests*, New York: Wiley, 1950) definitions of a purely speeded test and a purely power test, a new statistic called the "Speededness Quotient" (SQ) is proposed and defined as the percentage of unattempted items in the total number of errors.

The SQ is derived by dividing the total number of unattempted items by the total number of errors. The calculating formula is: $SQ = (NK - \Sigma A)/(N(K - M))$ where $N$ is the number of examinees; $K$ is the total number of items; $A$ is the number of items attempted by each individual, which includes the number of rights, omits, and wrongs; $M$ is the mean score.

The author states that the *SQ* gives a statistic that is invariant over the number of examinees and items when time is proportional. It is stressed that there is no such thing as the *SQ* of a test but only the *SQ* of a test with a given population. It is suggested that the test publishers should report the *SQ* of a test especially when reporting the reliability as an odd-even correlation coefficient.

Stallings, W. M. & Gillmore, G. M. A note on "accuracy" and "precision." *Journal of Educational Measurement*, 1971, *8*, 127-129.

In the literature of engineering and "hard" sciences, the term precision shares a common core meaning with reliability as used by behavioral scientists. Accuracy and validity have a similar semantic overlap.

In educational and psychological measurement, there is an interchangeable usage of accuracy and precision in defining reliability.

The authors of this paper advocate the use of precision, rather than accuracy, in describing reliability.

Stanley, J. C. & Wang, M. D. Weighting test items and test-item opinions, an overview of the analytic and empirical literature. *Educational and Psychological Measurement*, 1970, *30*, 21-35.

The authors conclude that differential weighting of a considerable number of positively intercorrelated item scores, with the weight for item *i* the same for all examinees, seems quite unpromising. They note that Birnbaum's (In Lord & Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968. 453-459) differential weighting of test items by levels of ability of examinees has produced interesting preliminary results that are likely to be followed up, and that Cleary's (*Psychometrika*, 1966, *31*, 215-224) procedure for securing regression weights that vary from examinee to examinee might improve the predictive validity of test scores. Criterion-keying of the options of right-answer items using Guttman's (In P. Horst (Ed.), *The prediction of personal adjustment*, New York: Social Science Research Council, 1941) procedure or perhaps a modification of deFinetti's (*British Journal of Mathematical & Statistical Psychology*, 1965, *18*, 87-123) approach to option marking via personal probabilities seems worth investigating.

Swaminthan, H. Hambleton, R. L. & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, *12*, 87-98.

The authors present an exposition of a decision-theoretic solution to the problem of allocating individuals to mastery states on the objectives included in a criterion-referenced test. Decisions are made by taking into account prior and collateral information on the examinees and also the losses associated with misclassifications.

Swaminathan, H., Hambleton, R. K. & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 263-267.

The purpose of this article was to describe a decision-theoretic formulation of criterion-referenced test reliability. It has been suggested that the primary purpose of criterion-referenced testing in objective-based instructional programs is to classify examinees into mastery states (masters or non-masters) on the objectives included in the test. The authors define the reliability of CR test scores in terms of consistency of the decision-making process across repeated administrations of the test. Specifically, reliability of a CR test is defined as a measure of agreement above chance expectation between the decisions made about examinee mastery states in repeated test administrations for each objective measured by the criterion-referenced test.

Coefficient Kappa ($k$) takes into account the measure of agreement expected by chance alone. $k = (p_0 - p_c)/(1 - p_c)$ where $p_0$ is the observed proportion of agreement and $p_c$ is the expected proportion of agreement. $k$ has an upper limit of +1 and a lower limit of close to −1. Since we usually have only a sample of examinees, $k$ must be estimated. $\hat{k}$ is defined as the sample analogue of $k = (p_0 - p_c)/(1 - p_c)$.

The authors conclude that the coefficient of agreement $k$ and hence the reliability of *CR* subtests is dependent on factors that affect the decision process. These factors include:

1) the method of assigning examinees to mastery states,
2) selection of the cutting score,
3) test length,
4) heterogeneity of the group.

The authors state that decision-making consistency is a measure of the reliability of the entire decision-making process, and that the test itself is only one input into the decision-making process. In generalizing reliability data to a new decision-making situation, all factors that affect the process must be considered.

Werts, C. E., Linn, R. L. & Jöreskog, K. A congeneric model for platonic true scores. *Educational and Psychological Measurement*, 1973, *33*, 311-318.

The authors provide an alternative formulation [to Levy (*Psychological Bulletin*, 1969, *71*, 276-277)] which allows for the model parameters to be determined given the structural specification of zero mean error and independence among errors for different items and between errors and true scores. Their approach is drawn from latent structure analysis (U. Grenander (Ed.), *Probability and statistics, the Harold Cramér volume.* New York: Wiley, 1959, 9-38) for the special case of dichotomous latent variables.

Werts, C. E., Linn, R. L. & Jöreskog, K. G. Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 1974, *34*, 25-33.

Intraclass correlation reliability estimates are based on the assumption that the various measures are equivalent. Jöreskog's (*Biometrika*, 1970, *57*, 239-251) general model for the analysis of covariance structures can be used to test the validity of this assumption.

Whitely, S. E. & Dawis, R. E. The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, 1974, *11*, 163-178.

Rasch and Wright have claimed that the Rasch model leads to a higher degree of objectivity in measurement than has been previously possible. Whitely and Dawis found that this is not so.

The authors conclude that the lack of impact of the Rasch model in test development is due more to the current status of trait measurement than to the properties of the model. Many of the advantages of the Rasch model necessitate a different kind of data for trait measurement than is now characteristic of the field. Explicit trait-item theory, locally independent items and routine administration of tests by computer, would be part of the necessary technological sophistication.

Wilbur, P. H. Positional response set among high school students on multiple-choice tests. *Journal of Educational Measurement*, 1970, *7*, 161-163.

The theory behind the manipulation of the position of the correct answer and of the most popular distractor was that if a positional response set should be present, having the correct answer in the position for which the individual is hypothesized to have a preference will tend to increase the score he would receive on those critical items.

Each test consisted of nine different forms in order to manipulate the critical items in nine different ways.

Form 1—had the correct answer for the 20 critical items in the first position,
Form 2—had the most popular distractor in the first position,
Form 3—had the correct answer in the second position,
Form 4—had the most popular distractor in the second position.
This procedure was followed for eight forms.

Form 9—had the correct answer randomized through the test with the restriction that each of the four response positions contained the correct answer exactly 20 times.

The results of this study lend support to the hypothesis that multiple-choice objective examinations are relatively free of examinee response bias. No significant evidence emerged for the existence of a universal positional response set in this study.

The author states that no consideration of intra-individual positional response set was made in this study and it may be that such a response bias might exist.

Wofford, J. C. The effects of item analysis methods and confidence levels upon test validity and cross-validity. *Journal of Educational Measurement*, 1968, *5*, 109-114.

The intent of this study is to investigate the effects of various methods of item selection and various confidence levels upon concurrent validity and cross-validity. Direct item-discrimination and item-total score analysis methods as well as validity maximizing methods are compared.

Results indicate that the cross-validities for this data are higher than the concurrent validities.

The use of item-discrimination and validity maximizing methods markedly increased the validity and cross-validity of the test over the unselected total test score. The item-total score methods did not yield tests of materially higher validity than was found for the total test.

A decided advantage was apparent for the use of validity maximizing methods in preference to the other methods. For situations in which high test validity is desired, one should choose the validity maximizing method.

Results also indicate that the lower confidence level generally is found to yield the more valid tests.

The author cautions that the results of this study should not be too broadly generalized.

Zimmerman, D. W. An item sampling model for the reliability of composite tests. *Educational and Psychological Measurement*, 1969, *29*, 49-59.

The author shows conditions under which KR-20, KR-21, and Guttman's lambda one are equal to reliability defined as a product moment correlation.

Zimmerman, D. W. Variability of test scores and the split-half reliability coefficient. *Educational and Psychological Measurement*, 1970, *30*, 259-266.

The purpose of this paper is to determine necessary and sufficient conditions under which the split-half reliability coefficient, defined with respect to a propensity distribution of half-test scores, is equal to the reliability of a test, defined with respect to a propensity distribution of total scores.

The results indicate that, whatever the reliability of a test may be, and whatever the source of variability in scores may be, the parameter value of the corrected split-half reliability coefficient based on random splits is given by KR-21. The present derivation extends Lyerly's (*Psychometrika*, 1958, *23*, 267-270) results (which proved that for a given set of observed scores the correlation between half-test scores over repeated splits and over persons, corrected by the Spearman-Brown formula, is given by KR-21) by indicating that the correlation between half-test scores over repeated splits, over persons, and over repeated testings resulting in different sets of observed scores, corrected by the Spearman-Brown formula, is also given by KR-21.

The author notes that the expected value of the sample KR-21 coefficient does not equal the expected value of the sample corrected split-half coefficient, nor does the expected value of either of these sample estimates equal the reliability of a test. Although necessary and sufficient conditions under which the parameter values of quantities such as KR-20, KR-21, and $\varrho_s$ (the corrected split-half reliability coefficient) are equal to $\varrho$ (reliability) can be stated, it is not known how departures from these conditions affect the bias and efficiency of sample estimates of reliability.

## BIBLIOGRAPHY

American Psychological Association. *Standards for educational and psychological tests and manuals.* Washington, D.C.: APA, 1966.

Bahadur, R. R. Examples of inconsistency of maximum likelihood estimates. *Sankhya*, 1958, *20*, 207-210.

Basu, D. An inconsistency of the method of maximum likelihood. *Annals of Mathematical Statistics*, 1955, *26*, 144-145.

Bennett, B. M. & Underwood, R. E. On McNemar's test for the 2x2 table and its power function. *Biometrics*, 1970, *26*, 339-343.

Berry, K. J., Martin, T. W., & Olson, K. F. A note on fourfold point correlation. *Educational and Psychological Measurement*, 1974, *34*, 53-56.

Bishop, Y., Fienberg, S. E., & Holland, P. W. *Discrete multivariate analysis.* Cambridge, Mass.: MIT Press, 1975.

Boschloo, R. D. Raised conditional level of significance for the 2x2 table when tested for the equality of two probabilities. *Statistica Neerlandica*, 1970, *21*, 1-35.

Cardinet, J., Tourneur, Y., & Allal, L. The symmetry of generalizability theory: Applications to educational measurements. *Journal of Educational Measurement*, 1976, *13*, 119-135.

Chandler, J. P. "STEPIT." *Behavioral Science*, 1969, *14*, 81-82.

Clopper, C. J., & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934, *26*, 404-413.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psycological Measurement*, 1960, *20*, 37-46.

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, *4*, 907-949.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley & Sons, 1972.

Fix, E. Tables of noncentral $\chi^2$. *University of California Publications in Statistics.* Berkeley, Vol. 1, No. 2, 1949.

Fleiss, J. L. *Statistical methods for rates and proportions.* New York: John Wiley & Sons, 1973.

Garside, G. R. An accurate correction for the $\chi^2$ test in the homogeneity case of the 2x2 contingency tables. *New Jersey Statistical Operations Research.* 1971, *7*, 1-26.

Garside, G. R., & Mack, C. Actual type 1 error probabilities for various tests in in the homogeneity case of the 2x2 contingency table. *The American Statistician*, 1976, *30*, 18-21.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross-classifications. *Journal of the American Statistical Association*, 1954, *49*, 732-764.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross-classifications II. *Journal of the American Statistical Association*, 1959, *54*, 123-163.

Hannan, J. Consistency of maximum likelihood estimation of discrete distributions. *Contributions to probability and statistics, essays in honor of Harold Hotelling.* Palo Alto: Stanford University Press, 1960.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement.* CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hayes, W. L. *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart & Winston, 1973.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation.* CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hogg, R. V., & Craig, A. T. *Introduction to mathematical statistics.* New York: Macmillan, 1970.

Horst, P. *Psychological measurement and prediction.* Belmont, Calif.: Wadsworth Publishing Co., 1966.

Johnson, N. L., & Kotz, S *Discrete distributions.* Boston: Houghton Mifflin, 1969.

Kaiser, H. F., & Michael, W. B. Domain validity and generalizability. *Educational and psychological measurement*, 1975, *35*, 31-35.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 2). New York: Hafner, 1961.

Kiefer, J., & Wolfowitz, J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 1956, *27*, 887-906.

Loeve, M. *Probability theory* (3rd ed.). Princeton, N. J.: Van Nostrand, 1963.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison Wesley, 1968.

Mantel, N., & Hankey, B. F. The odds ratio of a 2x2 contingency table. *American Statistician*, 1975, *29*, 143-145.

Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. *Educational and Psychological Measurement*, 1967, *27*, 335-348.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Neyman, J., & Scott, E. L. Consistent estimates based on partially consistent observations. *Econometrika*, 1948, *16*, 1-32.

Rao, C. R. *Linear statistical inference and its applications.* New York: John Wiley & Sons, 1973.

Rudin, W. *Principles of mathematical analysis.* New York: McGraw Hill, 1964.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 1955, *19*, 321-325.

Shoemaker, D. M. Toward a framework for achievement testing. *Review of Educational Research*, 1975, *45*, 127-147.

Sirotnik, K. An analysis of variance framework for matrix sampling. *Educational and Psychological Measurement*, 1970, *30*, 891-908.

Walker, H. M., & Lev, J. *Statistical inference.* New York: Henry Holt & Co., 1953.

Wilcox, R. R. The stability and equivalence of achievement test items. Unpublished doctoral dissertation, University of California, Santa Barbara, 1976.

Wilks, S. S. *Mathematical statistics.* New York: John Wiley & Sons, 1962.

Zacks, S. The theory of statistical inference. New York: John Wiley & Sons, 1971.

Zehna, P. W. Invariance of maximum likelihood estimation. *Annals of Mathematical Statistics*, 1966, *37*, 744.

## ADDITIONAL REFERENCES

Airasian, P. W. Formative evaluation instruments: A construction and validation of tests to evaluate learning over short time periods. Unpublished doctoral dissertation, University of Chicago, 1969.

Airasian, P. W., & Bart, W. Tree Theory: A theory-generative measurement model. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Airasian, P. W., & Madaus, G. F. Criterion-referenced testing in the classroom. *Measurement in Education*, 1972, *3*, 1-8.

Alexander, H. W. The estimation of reliability when several trials are available. *Psychometrika*, 1947, *12*, 79-99.

Alkin, M. C. Criterion-referenced measurement and other such terms. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement.* CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Anatasi, A., & Drake, J. D. An empirical comparison of certain techniques for estimating the reliability of speeded tests. *Educational and Psychological Measurement*, 1954, *14*, 529-540.

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. *The British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 231-238.

Anderson, T. W. Some scaling models and estimation procedures in the latent class model In U. Grenander (Ed.), *Probability and statistics, the Harold Cramer volume.* New York: John Wiley & Sons, 1959.

Angoff, W. H. Test reliability and effective test length. *Psychometrika*, 1953, *18*, 1-14.

Appel, V., & Kipnis, D. The use of levels of confidence in item analysis. *Journal of Applied Psychology*, 1954, *38*, 256-259.

Archer, N. W. Effects of confidence weighting by fifth and sixth grade students on objective test scores. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1963.

Arnold, J. C., & Arnold, P. L. On scoring multiple choice exams allowing for partial knowledge. *The Journal of Experimental Education*, 1970, *39*, 8-13.

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology*, 1974, *14*, 10-16.

Baker, F. B. An intersection of test score interpretation and item analysis. *Journal of Educational Measurement*, 1964, *1*, 23-28.

Baker, F. B. Computer-based instructional management systems: A first look. *Review of Educational Research*, 1971, *41*, 51-70.

Baker, F. B. & Hoyt, C. J. The relation of the method of reciprocal averages to Guttman's internal consistency scaling model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Barcikowski, R. S. Optimum use of the item sampling technique in obtaining test norms. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Barcikowski, R. S., & Terranova, C. Item sampling. Paper presented at the annual meeting of the Educational Research Association of New York State, 1966.

Beck, I. L., & Mitroff, D *Rationale and design of a primary grades reading system for an individualized classroom*. Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1972.

Bendig, A. W. Reliability and the number of rating scale categories. *Journal of Applied Psychology*, 1954, *38*, 38-40.

Berkson, J. Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association*, 1955, *50*, 13-162.

Block, J. H. Criterion-referenced measurements: Potential. *School Review*, 1971, *69*, 289-298.

Block, J. H. *Mastery learning: Theory and practice*. New York: Holt, Rinehart, & Winston, 1971.

Block, J. H. Student learning and the setting of mastery performance standards. *Educational Horizons*, 1972, *50*, 183-190.

Bloom, B. S. Learning for mastery. *Evaluation Comment*, 1968, *1*(2).

Bloom, B. S. (Ed.) *Taxonomy of educational objectives: Handbook I, cognitive domain*. New York: Longmans, Green & Co., 1956.

Bormuth, J. R. Cloze as a measure of readability. *Yearbook of the International Reading Association*. Newark, Delaware: International Reading Association, 1963, *8*, 131-134.

Bormuth, J. R. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 1967, *10*, 291-299.

Bormuth, J. R. Development of standards of readability: Toward a rational criterion of passage performance. Final report, USDHEW, project No. 9-0237. Chicago: The University of Chicago, 1971.

Bormuth, J. R. *On the theory of achievement test items*. Chicago, University of Chicago Press, 1970.

Boruch, R. F., & Wolins, L. A. A procedure for estimation of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement*, 1970, *30*, 547-574.

Brennan, R. L. Some statistical problems in the evaluation of self-instructional programs. Unpublished doctoral dissertation, Harvard University, 1970.

Brennan, R. L. The evaluation of mastery test items, U. S. Office of Education, project No. 2B118, 1974.

Brennan, R. L., & Kane, M. T. The generalizability of class means. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., 1975.

Brennan, R. L., & Stolurow, L. M. An empirical decision process for formative evaluation. *Research Memorandum No.4*. Harvard CAI Laboratory, Cambridge, Mass., 1971.

Brenner, M. H. Test difficulty, reliability, and discrimination as functions of item difficulty order. *Journal of Applied Psychology*, 1964, *48*, 98-100.

Brogden, H. E. Effect of bias due to difficulty factor in product-moment item intercorrelations on the accuracy of estimation of reliability by the Kuder-Richardson Formula No. 20. *Educational and Psychological Measurement*, 1946, *6*, 517-520.

Brogden, H. E. Variation in test validity with variation in the distribution of item difficulties, numbers of items, and degrees of their intercorrelation. *Psychometrika*, 1946, *11*, 197-214.

Brooks, R. D. An empirical investigation of the Rasch ratio-scale model for item difficulty indexes. Unpublished doctoral dissertation, University of Iowa, 1965.

Burmester, M. A., & Olson, L. A. Comparison of item statistics for items in multiple choice and alternate response form. *Science Education*, 1966, *50*, 467-470.

Burt, C. Test reliability estimated by analysis of variance. *The British Journal of Statistical Psychology*, 1955, *8*, 103-118.

Burton, N. W., & Remer, R. An empirical application of the item sampling technique to questionnaire data. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1972.

Carroll, J. B. A model of school learning. *Teachers College Record*, 1963, *64*, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. *Educational Horizons*, 1970, *48*, 71-80.

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 1945, *10*, 1-19.

Carver, R. P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 1974, *29*, 512-518.

Cashen, V. M., & Ramseyer, G. C. The use of separate answer sheets by primary age children. *Journal of Educational Measurement*, 1969, *6*, 155-158.

Cochran, W. G. Errors of measurement in statistics. *Technometrics*, 1968, *10*, 637-666.

Cohen, J. Weighted chi-square: An extension of the kappa method. *Educational and Psychological Measurement*, 1972, *32*, 61-74.

Cohen, J. Weighted kappa: Nominal scale agreement with provisions for scaled disagreement of partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Cook, D. L., & Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. *Journal of Educational Measurement*, 1967, *4*, 27-33.

Cooley, W. W., & Glaser, R. An information system for individually prescribed instruction. In R. C. Atkinson & H. A. Wilson (Eds.), *Computer assisted instruction*. New York: Academic Press, 1969.

Cooley, W. W., & Glaser, R. The computer and individualized instruction. *Science*, 1969, *166*, 574-582.

Committee on Student Appraisal & Office of Research in Medical Education. *A taxonomy of intellectual processes*. Chicago: University of Illinois College of Medicine, 1962.

Coombs, C. H. On the use of objective examinations. *Educational and Psychological Measurement*, 1953, *13*, 308-310.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial knowledge. *Educational and Psychological Measurement.*, 1956, *16*, 13-37.

Coulson, D. B., & Hambleton, R. K. On the validation of criterion-referenced tests designed to measure individual mastery. Paper presented at the annual meeting of the American Psychological Association, New Orleans, 1974.

Cox, R. C. Evaluation aspects of criterion-referenced measurement. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Cox, R. C., & Boston, M. E. Diagnosis of pupil achievement in the individually prescribed instruction project. Working Paper 15. Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1967.

Cox, R. C., & Graham, G. T. The development of a sequentially scaled achievement test. *Journal of Educational Measurement*, 1966 *3*, 147-150.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

Cronbach, L. J. Course improvement through evaluation. *Teachers College Record*, 1963, *64*, 672-683.

Cronbach, L. J. Further evidence on response sets and test design. *Educational and Psychological Measurement*, 1950, *10*, 3-31.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles E. Merrill, 1967.

Cronbach, L. J. Review of *On the theory of achievement test items* by J. R. Bormuth, with an appendix by P. Menzel. *Psychometrika*, 1970, *35*, 509-511.

Cronbach, L. J. Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, 1942, *33*, 401-415.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Cronbach, L. J., & Azuma, H. Internal-consistency reliability formulas applied to randomly sampled single-factor tests: An empirical comparison. *Educational and Psychological Measurement*, 1965, *25*, 291-321.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* Urbana: University of Illinois Press, 1965.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281-302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, *16*, 137-163.

Cronbach, L. J., Schönemann, R., & McKie, D. Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 1965, *25*, 291-321.

Cronbach, L. J., & Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika*, 1952, *17*, 127-147.

Cronbach, L. J., & Warrington, W. G. Time limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, 1951, *16*, 167-188.

Culhane, T. Q., & Stodola, Q. C. Use of mark-sense cards with elementary school children. *Educational and Psychological Measurement*, 1967, *27*, 183-185.

Cureton, E. E. Note on $\phi/\phi$ max. *Psychometrika*, 1959, *24*, 89-91.

Cureton, E. E. Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement*, 1965, *25*, 327-346.

Cureton. E. E. The correction for guessing. *Journal of Experimental Education*, 1966, *34*, 44-47.

Das, R. S. Item analysis by probit and fractile graphical methods. *British Journal of Statistical Psychology*, 1964, *27*, 51-64.

Davis, F. B. Estimation and use of scoring weights for each choice in multiple choice test items. *Educational and Psychological Measurement*, 1959, *19*, 291-298.

Davis, F. B. Item-analysis data: Their computation, interpretation, and use in test construction. *Harvard Educational Papers*, 1949, *2*, 1-42.

Davis, F. B. Item analysis in relation to educational and psychological testing. *Psychological Bulletin*, 1952, *49*, 97-121.

Davis, F. B. Item selection techniques. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D. C.: American Council on Education, 1951.

Davis, F. B. Notes on test construction: The reliability of item-analysis data. *Journal of Educational Psychology*, 1946, *37*, 385-390.

Davis, F. B. Use of correction for chance success in test scoring. *Journal of Educational Research*, 1959, *52*, 279-280.

Davis, F. B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, *19*, 159-170.

De Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, *18*, 87-123.

Diamond, J. J., & Evans, W. B. The correction for guessing. *Review of Educational Research*, 1973, *43*, 181-191.

Diederich, P. B. Review of *On the theory of achievement test items* by J. R. Bormuth, with an appendix by P. Menzel. *Educational and Psychological Measurement*, 1970, *30*, 1003-1005.

Donlon, T. F. Some needs for clearer terminology in criterion-referenced testing. Paper presented at the annual meeting of the Nationawl Council on Measurement in Education, Chicago, 1974.

Dunn, T. F., & Goldstein, L. G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, 1959, *19*, 171-179.

Dunn, O. J., & Clark, V. Comparison of tests of the equality of dependent correlation coefficients. *Journal of American Statistical Association*, 1971, *66*, 904-908.

Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 15-25.

Ebel, R. L. Criterion-referenced measurements: Limitations. *School Review*, 1971, *79*, 282-288.

Ebel, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, *16*, 407-424.

Ebel, R. L. Evaluation and educational objectives. *Journal of Educational Measurement*, 1973, *10*, 273-279.

Ebel, R. L. Expected reliability as a function of choices per item. *Educational and Psychological Measurement*, 1969, *29*, 565-570.

Ebel, R. L. Must all tests be valid? *American Psychologist*, 1961, *16*, 640-647.

Ebel, R. L. Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 1954, *14*, 352-353.

Ebel, R. L. The case for true-false items. *School Review*, 1970, *78*, 373-389.

Ebel, R. L. The comparative effectiveness of true-false and multiple choice achievement test items. Paper presented at the annual meeting of the American Educational Research Association, New York City, 1971.

Echternacht, G. J. The use of confidence testing in objective tests. *Review of Educational Research*, 1972, *42*, 217-236.

Ely, J. H. An empirical evaluation of the effects of various methods of item analysis upon test reliability. Unpublished doctoral dissertation, Purdue University, 1950.

Englehart, M. D. A comparison of several item discrimination indices. *Journal of Educational Measurement*, 1965, *2*, 69-76.

Englehart, M. D. A method of estimating the reliability of ratings compared with certain methods of estimating the reliability of tests. *Educational and Psychological Measurement*, 1959, *19*, 479-588.

Everitt, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 97-103.

Feldman, M. J. The effects of the size of criterion group and the level of significance in selecting test items on the validity of tests. *Educational and Psychological Measurement*, 1953, *13*, 273-279.

Feldt, L. S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, 1969, *34*, 363-373.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 1965, *30*, 357-370.

Feldt, L. S., & Forsyth, R. A. An examination of the context effect in item sampling. *Journal of Educational Measurement*, 1974, *11*, 73-82.

Fhaner, S. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 172-175.

Findley, W. G. A rationale for evaluation of item discrimination statistics. *Educational and Psychological Measurement*, 1956, *16*, 175-180.

Finney, D. J. The application of probit analysis to the results on mental tests. *Psychometrika*, 1944, *19*, 31-39.

Flanagan, J. C. Discussion of symposium: Standard scores for aptitude and achievement tests. *Educational and Psychological Measurement*, 1962, *22*, 35-39.

Flanagan, J. C. Units, scores and norms. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, D. C.: American Council on Education, 1951.

Fleiss, J. L., Cohen, J., & Everitt. B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Frary, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. *Educational and Psychological Measurement*, 1969, *29*, 359-365.

Fremer, J., & Anastasio, E. J. Computer-assisted item writing—I (Spelling items). *Journal of Educational Measurement*, 1969, *6*, 69-74.

Frisbie, D. A. Comparative reliabilities and validities of true-false and multiple choice tests. Unpublished doctoral dissertation, Michigan State University, 1971.

Gaylord, R. H. Estimating test reliability from the item-test correlations. *Educational and Psychological Measurement*, 1969, *29*, 303-304.

Glaser, R. L. Instructional Technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 1963, *18*, 519-521.

Glaser, R. L., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thordike (Ed.), *Educational measurement*. Washington, D. C.: American Council on Education, 1971.

Glass, G. V. A ranking variable analogue of biserial correlation: Implications for short-cut item analysis. *Journal of Educational Measurement*, 1965, *2*, 91-95.

Glass, G. V. Note on rank biserial correlation. *Educational and Psychological Measurement*, 1966, *26*, 623-631.

Glass, G. V., & Stanley, J. C. Effects of correction for differential omissions on the internal consistency of tests. Paper presented at the annual meeting of the Psychonomic Society, St. Louis, 1962.

Gleser, G. C., Cronbach, J. J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 1965, *30*, 395-418.

Gleser, G. C., & DuBois, P. H. A successive approximation method of maximizing test validity. *Psychometrika*, 1951, *16*, 129-139.

Gruber, H. E., & Weitman, M. Item analysis and the measurement of change. *Journal of Educational Research*, 1962, *6*, 287-289.

Gulliksen, H. Intrinsic validity. *American Psychologist*, 1950, *5*, 511-517.

Gulliksen, H. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika*, 1945, *10*, 79-91.

Gulliksen, H. The reliability of speeded tests. *Psychometrika*, 1950, *15*, 259-269.

Gulliksen, H. *Theory of mental tests.* New York: John Wiley & Sons, 1950.

Gulliksen, H., & Wilks, G. S. Regression tests for several samples. *Psychometrika*, 1950, *15*, 91-104.

Gumbel, E. J. Bivariate logistic distributions. *Journal of the American Statistical Association*, 1961, *56*, 335-349.

Gurland, J., Ilbok, L., & Dahm, P. A. Polychotomous quantal response in biological assay. *Biometrics*, 1960, *16*, 382-398.

Gustav, A. Response set in objective achievement tests. *Journal of Psychology*, 1963, *56*, 421-427.

Guttman, L. A basis for analyzing test-retest reliability. *Psychometrika*, 1945, *10*, 255-282.

Guttman, L. A basis for scaling qualitative ideas. *American Sociological Review*, 1944, *9*, 139-150.

Guttman, L. A general non-metric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 1968, *33*, 469-506.

Guttman, L. A new approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, Illinois: Free Press, 1954.

Guttman, L. Supplementary study B. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.

Guttman, L. The basis for scalagram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lasarsfeld, S. A. Star, & J. A. Clausen, *Measurement and prediction.* Princeton, N. J.: Princeton University Press, 1950.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst (Ed.), *The prediction of personal adjustment.* New York: Social Science Research Council, 1941.

Guttman, L. The test-retest reliability of qualitative data. *Psychometrika*, 1946, *11*, 81-95.

Guttman, L., & Schlesinger, I. M. Systematic construction of distractors for ability and achievement test items. *Educational and Psychological Measurement*, 1967, *27*, 159-170.

Haggard, E. *Intraclass correlation and the analysis of variance.* New York: Dryden Press, 1958.

Hambleton, R. K. A review of testing and decision-making procedures for selected individualized instructional programs. *ACT Technical Bulletin*, No. 15. Iowa City, Iowa: The American College Testing Program, 1973.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research*, 1974, *44*, 371-400.

Handler, H. The instructional objectives controversy. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1968.

Hanna, G. S. Improving reliability and validity of multiple-choice tests with an answer-until-correct procedure. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1974.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. *Sample survey methods and theory* (Vol. 1). New York: John Wiley & Sons, 1953.

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 1972, *9*, 27-30.

Harris, C. W. Note on the variances and covariances of three error types. *Journal of Educational Measurement*, 1973, *10*, 49-50.

Harris, C. W. Problems of objectives-based measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Harris, C. W., Alkin, M. C., & Popham, W. J. (Eds.), *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Harris, M. L., & Stewart, D. M. Application of classical strategies to criterion-referenced test construction. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hayward, P. A comparison of test performance on three answer sheet formats. *Educational and Psychological Measurement*, 1967, *27*, 997-1004.

Heathers, G. Overview of innovations in organization for learning. *Interchange*, 1972, *3*, 47-68.

Helmstadter, G. C. Comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance-oriented instruction. Paper presented at the annual meeting of the American Psychological Association, Hawaii, 1972.

Hendrickson, G. F. The effect of differential option weighting on multiple-choice objective tests. Report No. 93. The Center for Social Organization of Schools, Johns Hopkins University, 1971.

Hendrickson, G. F., & Green, B. F. Comparison of the factor structure of Guttman-weighted vs. rights only weighted tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Henrysson, S. Different measures of reliability. *Research Bulletin*. Stockholm, Sweden: Institute of Educational Research, Teachers College, 1959.

Henrysson, S., & Wedman, I. Some problems in construction and evaluation of criterion-referenced tests. *Scandinavian Journal of Educational Research*, 1974, *18*, 1-12.

Hieronymus, A. N. Today's testing: What do we know how to do? In *Proceedings of the 1971 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1972.

Hoffman, B. *The tyranny of testing*. New York: Collier, 1964.

Hooke, R. Symmetric functions of a two-way array. *Annals of Mathematical Statistics*, 1956, *27*, 55-79.

Hopkins, K. D. Extrinsic reliability: Estimating and attenuating variance from response styles, chance, and other irrelevant sources. *Educational and Psychological Measurement*, 1964, *24*, 271-281.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement*, 1973, *33*, 135-141.

Hopkins, K. D., & Hopkins, B. R. Intraindividual and interindividual positional preference response styles in ability tests. *Educational and Psychological Measurement*, 1964, *24*, 801-805.

Horn, J. L. Equations representing combinations of components in scoring psychological variables. *Acta Psychologica*, 1963, *21*, 184-217.

Horst, P. Item selection by means of a maximizing function. *Psychometrika*, 1936, *1*, 229-244.

Horst, P. The chance element in the multiple-choice test item. *Journal of General Psychology*, 1933, *24*, 229-232.

Hoyt, C. Test reliability estimated by analysis of variance. *Psychometrika*, 1941, *6* 153-160.

Hsu, T. Empirical data on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hurst, J., Bartlett, J., & Roming, R. An approach to evaluation in educational psychology courses and its instrumentation. *Educational and Psychological Measurement*, 1961, *21*, 445-456.

Husek, T. R., & Sirotnik, K. Item sampling in educational research: An empirical investigation. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1968.

Ivens, S. H. An investigation of item analysis, reliability and validity in relation to criterion-referenced tests. Unpublished doctoral dissertation, Florida State University, 1970.

Jackson, P. H. Simple approximations in the estimation of many parameters. *British Journal of Mathematical and Statistical Psychology*, 1972, *25*, 213-229.

Jackson, P. H., & Novick, M. R. Maximizing the validity of a unit-weight composite as a function of relative component lengths with a fixed total testing time. *Psychometrika*, 1970, *35*, 333-347.

Jackson, R. W. B., & Ferguson, G. A. Studies on the reliability of tests. Bulletin No. 12. Toronto: University of Toronto, 1941.

Jacobs, P. I., & Vandeventer, M. Information in wrong responses. Research Bulletin 68-25. Princeton, N. J.: Educational Testing Service, 1968.

Johnson, A. P. Notes on a suggested index of item validity: The U-L index. *Journal of Educational Psychology*, 1951, *42*, 499-504.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 373-380.

Jöreskog, K. G. A general method for analysis of covariance structures. *Biometrika*, 1970, *57*, 239-251.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, *36*, 109-134.

Jöreskog, K. G., van Thillo, M., & Gruaevus, G. T. ACOVSM—a general computer program for analysis of covariance structures including generalized MANOVA. Research Bulletin 71-1. Princeton, N. J.: Educational Testing Service, 1971.

Kaiser, H. F. A second generation "Little Jiffy." *Psychometrika*, 1970, *35*, 401-415.

Kaiser, H. F. Scaling a simplex. *Psychometrika*, 1962, *27*, 255-262.

Kaiser, H. F. Uncorrelated linear composites maximally related to a complex of correlated observations. *Educational and Psychological Measurement*, 1967, *27*, 3-6.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. *Psychometrika*, 1962, *27*, 59-72.

Kelley, T. L. *Statistical method.* New York: Macmillan Co., 1923.

Kelley, T. L. The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 1939, *30*, 17-24.

Kendall, M. G. *Rank correlation methods* (3rd ed.). London: Griffin, 1962.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics*, (Vol. 1). New York: Hafner, 1963.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics*, (Vol. 3). New York: Hafner, 1966.

Kirk, R. E. Experimental design: *Procedures for the behavioral sciences.* Belmont, California: Brooks & Cole, 1968.

Klein, D. F., & Cleary, T. A. Platonic true scores and error in psychiatric rating scales. *Psychological Bulletin*, 1967, *68*, 77-80.

Klein, D. F., & Cleary, T. A. Platonic true scores: Further comment. *Psychological Bulletin*, 1969, *71*, 278-280.

Knapp, T. R. An application of balanced incomplete block design to the estimation of test norms. *Educational and Psychological Measurement*, 1968, *28*, 265-272.

Komorita, S. S., & Graham, W. K. Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, 1965, *25*, 987-995.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Kristof, W. Testing differences between reliability coefficients. *The British Journal of Statistical Psychology*, 1964, *17*, 105-111.

Kristof, W. The statistical theory of stepped-up reliability coefficients when a test nas been divided into several equivalent parts. *Psychometrika*, 1963, *28*, 221-238.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151-160.

Lawley, D. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, 1943, 61-A, Part 3, 273-287.

Lehmann, E. L. *Testing statistical hypotheses.* New York: John Wiley & Sons, 1959.

Levy, P. Platonic true scores and rating scales: A case of uncorrelated definitions.

*Psychological Bulletin*, 1969, *71*, 276-277.

Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. *ACT Technical Bulletin* No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Light, R. J. Issues in the analysis of qualitative data. In R. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973.

Lindquist, E. F. Basic considerations in answer sheet design. *Testing Today*. Houghton Mifflin circular, 1964.

Lindquist, E. F. *Design and analysis of experiments*. Boston: Houghton Mifflin, 1953.

Lindquist, E. F. (Ed.) *Educational Measurement*. Washington, D. C.: American Council on Education, 1951.

Lindvall, C. M., & Cox, R. The role of evaluation in programs for individualized instruction. In R. W. Tyler (Ed.), *Educational evaluation: New roles, new means*. Sixty-eighth Yearbook, Part II. Chicago: National Society for the Study of Education, 1969.

Lindvall, C. M., Cox, R. C., & Bolvin, J. O. *Evaluation as a tool in curriculum development: The IPI evaluation program*. AERA monograph series on curriculum evaluation, No. 5. Chicago: Rand McNally, 1970.

Livingston, S. A. A note on the interpretation of the criterion-referenced reliability coefficient. *Journal of Educational Measurement*, 1973, *10*, 311.

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." *Journal of Educational Measurement*, 1972, *9*, 31.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26.

Loadman, W. E. An inquiry concerning the use of item sampling as a method to reduce testing time in an evaluation study. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1972.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, *3*, 635-694.

Loevinger, J. Person and population as psychometric concepts. *Psychological Review*, 1965, *72*, 143-155.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. An approach to mental test theory. *Psychometrika*, 1959, *24*, 282-302.

Lord, F. M. An empirical study of item-test regression. *Psychometrika*, 1965, *30*, 373-376.

Lord, F. M. A note on the normal ogive or logistic curve in item analysis. *Psychometrika*, 1965, *30*, 371-372.

Lord, F. M. A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika*, 1957, *22*, 207-220.

Lord, F. M. A theoretical study of the measurement effectiveness of flexi-level tests. *Educational and Psychological Measurement*, 1971, *31*, 805-813.

Lord, F. M. A theory of test scores. *Psychometric Monographs*, No. 7, 1952.

Lord, F. M. Do tests of the same length have the same standard errors of measurement? *Educational and Psychological Measurement*, 1957, *17*, 510-521.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change.* Madison, Wisconsin: University of Wisconsin Press, 1963.

Lord, F. M. Equating test scores—a maximum likelihood solution. *Psychometrika*, 1955, *20*, 193-200.

Lord, F. M. Estimating test reliability. *Educational and Psychological Measurement*, 1955, *15*, 325-336.

Lord, F. M. Estimating norms by item-sampling. *Educational and Psychological Measurement*, 1962, *22*, 259-267.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). *Psychometrika*, 1969, *34*, 259-299.

Lord, F. M. Formula scoring and validity. *Educational and Psychological Measurement*, 1963, *23*, 663-672.

Lord, F. M. Guessing and test performance. *Educational and Psychological Measurement*, 1963, *23*, 663-672.

Lord, F. M. Item sampling in test theory and in research design. Research Bulletin RB-65-22. Princeton, N. J.: Educational Testing Service, 1965.

Lord, F. M. Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, 1971, *31*, 3-31.

Lord, F. M. Sampling fluctuations resulting from the sample of test items. *Psychometrika*, 1955, *20*, 1-22.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing and guidance.* New York: Harper and Row, 1971.

Lord, F. M. Statistical inferences about true scores. *Psychometrika*, 1959, *24*, 1-18.

Lord, F. M. Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, 1959, *19*, 233-239.

Lord, F. M. Test reliability—a correction. *Educational and Psychological Measurement*, 1962, *22*, 511-512.

Lord, F. M. The relations of test score to the trait underlying the test. *Educational and Psychological Measurement*, 1953, *4*, 517-549.

Lord, F. M. The relative efficiency of two tests as a function of ability level. *Psychometrika*, 1974, *39*, 351-358.

Lord, F. M. Use of the true-score theory to predict moments of unvariate and bivariate observed-score distributions. *Psychometrika*, 1960, *25*, 325-342.

Loveland, E. H. Measurement of factors affecting test-retest reliability. Unpublished doctoral dissertation, University of Tennessee, 1952.

Lu, K. H. A measure of agreement among subjective judgments. *Educational and Psychological Measurement*, 1971, *31*, 75-84.

Lyerly, S. B. A note on correcting for chance successes in objective tests. *Psychometrika*, 1951, *16*, 21-30.

Lyerly, S. B. The Kuder-Richardson formula (21) as a split half coefficient and some remarks on its basic assumption. *Psychometrika*, 1958, *23*, 267-270.

Mager, R. F. *Preparing instructional objectives.* Palo Alto: Fearon Publishers, 1962.

Magnusson, D. *Test theory*. Reading, Mass.: Addison-Wesley, 1967.

Marcus, A. The effect of correct response location in the difficulty level of multiple-choice questions. *Journal of Applied Psychology*, 1963, *47*, 48-51.

Matell, M. S., & Jacoby, J. Is there an optimal number of alternatives for Likert scale items? Study 1: Reliability and validity. *Educational and Psychological Measurement*, 1971, *31*, 657-674.

Maxwell, A. E. Maximum likelihood estimates of item parameters using the logistic function. *Psychometrika*, 1959, *24*, 221-227.

Maxwell, A. E., & Pilliner, A. E. G. Deriving coefficients of agreement for ratings. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 105-116.

Mayo, S. T. Mastery learning and mastery testing. *NCME Measurement in Education*, 1970, *3*, 1-4.

McDonald, R. P. A unified treatment of the weighting problem. *Psychometrika*, 1968, *33*, 351-381.

McGuire, C. H., & Babbott, D. Simulation technique in the measurement of problem-solving skills. *Journal of Educational Measurement*, 1967, *4*, 1-10.

McNamara, W. J., & Weitzman, E. The effect of choice placement on the difficulty of multiple-choice questions. *Journal of Educational Psychology*, 1945, *36*, 103-113.

Merwin, J. D. Rational and mathematical relationships of six scoring procedures applicable to 3-choice items. *Journal of Educational Psychology*, 1959, *50*, 153-161.

Michael, J. J. The reliability of multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, 1968, *5*, 307-314.

Michael, W. B., Stewart, R., Douglas, B., & Rainwater, J. H. An experimental determination of the optimal scoring formula for highly speeded tests under different instructions regarding scoring penalties. *Educational and Psychological Measurement*, 1963, *23*, 83-99.

Miklich, D. R., & Gordon, G. Test-taking carefulness vs. acquiescence response set on true-false examinations. *Educational and Psychological Measurement*, 1968, *28*, 545-548.

Miller, R. B. Task description and analysis. In R. Gagné (Ed.), *Psychological principles in system development*. New York: Holt, Rinehart & Winston, 1962.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current practices. Berkeley, California: McCutchan Publishers, 1974.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. *Phi Delta Kappan*, 1970, *52*, 226-230.

Millman, J., Bishop, H., & Ebel, R. An analysis of test-wiseness. *Educational and Psychological Measurement*, 1965, *25*, 707-726.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement*, 1974, *11*, 137-138.

Moonan, W. J. An item response generator. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1965.

Morrison, E. J. On test variance and the dimensions of the measurement situation. *Educational and Psychological Measurement*, 1960, *20*, 231-250.

Mosier, C. I. Machine methods in scaling by reciprocal averages. Proceedings, *Research Forum*, New York: International Business Machines Corporation, 1946.

Mosier, C. I. On the reliability of a weighted composite. *Psychometrika*, 1943, *8*, 161-168.

Munz, D. C., & Smouse, A. D. The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology*, 1968, *68*, 181-184.

Munz, D. C., & Smouse, A. D. Interaction effects of item difficulty sequence and achievement-anxiety reaction on academic performance. *Journal of Educational Psychology*, 1968, *59*, 370-374.

Myers, C. T. Symposium: The effects of time limits on test scores. *Educational Measurement*, 1960, *20*, 221-222.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, 1954, *14*, 459-472.

Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Nitko, A. J. The power functions of some proposed tests of the significance of coefficient alpha in the one-sample and two-sample cases. Unpublished doctoral dissertation, University of Iowa, 1968.

Novick, M. R. The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 1966, *6*, 1-18.

Novick, M. R., & Lewis, C. Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 1967, *32*, 1-13.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. *Psychometrika*, 1973, *38*, 19-46.

Olkin, I. Correlations revisited. In J. C. Stanley (Ed.), *Improving experimental design and statistical analysis.* Chicago: Rand McNally, 1967.

Owens, T. R., & Stufflebeam, D. L. An experimental comparison of item sampling and examinee sampling for estimating test norms. *Educational and Psychological Measurement*, 1969, *6*, 75-83.

Patnaik, D., & Traub, R. E. Differential weighting by judged degree of correctness. *Journal of Educational Measurement*, 1973, *10*, 281-286.

Payne, W. H., & Anderson, D. E. Significance levels for the Kuder-Richardson twenty: An automated sampling experiment approach. *Educational and Psychological Measurement*, 1968, *28*, 23-39.

Penfield, D. A. An empirical investigation of the approximate sampling distribution for Kuder-Richardson twenty. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1968.

Piper, R. M. Multiple-matrix sampling in the evaluation of a music program. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1972.

Plumlee, L. B. Estimating means and standard deviations from partial data—an empirical check on Lord's sampling technique. *Educational and Psychological Measurement*, 1964, *24*, 623-630.

Plumlee, L. B. The effect of difficulty and chance successes on item-test correlation and on test reliability. *Psychometrika*, 1945, *10*, 1-19.

Popham, W. J. (Ed.). *Criterion-referenced measurement: An Introduction.* Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Popham, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction.* Englewood Cliffs, N. J.: Educational Technology Publications, 1971.

Popham, W. J. Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Rabehl, G. J. The MINNEMAST experiment with domain referenced achievement testing. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Rahmlow, H. F., Matthews, J. J., & Jung, S. M. An empirical investigation of item analysis in criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, 1970.

Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. *Psychometrika*, 1960, *25*, 261-271.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. Generalizability of stratified-parallel tests. *Psychometrika*, 1965, *30*, 39-56.

Ramos, R. A., & Stern, J. Item behavior associated with changes in the number of alternatives in multiple choice items. *Journal of Educational Measurement*, 1973, *10*, 305-310.

Ramsay, J. O. A scoring system for multiple choice test items. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 247-250.

Rapaport, G. M., & Berg, I. A. Response bias in an unstructured questionnaire. *Journal of Psychology*, 1954, *37*, 475-481.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science*. Chicago: Science Research Associates, 1966.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, *19*, 49-57.

Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics*. Berkeley: University of California Press, 1961, IV, 321-334.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

Reilly, R. R., & Dynarski, B. J. A computer program for keying options of multiple-choice tests to increase internal consistency. *Educational and Psychological Measurement*, 1972, *32*, 789-791.

Richardson, M. W. Notes on the rationale of item analysis. *Psychometrika*, 1936, *1*, 69-76.

Richardson, M. W. The combination of measures. In P. Horst (Ed.), *The prediction of personal adjustment*. New York: Social Science Research Council, 1941.

Richardson, M. W. The relation between difficulty and the differential validity of a test. *Psychometrika*, 1936, *1*, 33-49.

Richardson, M. W., & Adkins, D. C. A rapid method of selecting test items. *Journal of Educational Psychology*, 1938, *29*, 547-552.

Rippey, R. A Fortran program for scoring and analyzing probabilistic tests. *Behavioral Science*, 1968, *13*, 424.

Rippey, R. Probabilistic testing. *Journal of Educational Measurement*, 1968, *5*, 211-215.

Robinson, W. S. The statistical measure of agreement. *American Sociological Review*, 1957, *22*, 17-25.

Ruch, G. M., & DeGraff, M. H. Correction for chance and 'guess' versus 'do not guess' instructions in multiple-response tests. *Journal of Educational Psychology*, 1926, *17*, 368-375.

Ruch, G. M., & Stoddard, G. D. Comparative reliabilities of five types of objective examinations. *Journal of Educational Psychology*, 1925, *16*, 89-103.

Sabers, D. L., & White, G. W. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement*, 1969, *6* 93-96.

Safford, J. Defining item forms for use with instructional objectives. Technical Paper No. 2, Instructional Objectives Exchange, Los Angeles, 1970.

Saupe, J. L. Selecting items to measure change. *Journal of Educational Measurement*, 1966, *3*, 223-228.

Sax, G., & Collet, L. An analysis and empirical study of the effects of guessing formulas and instructions on reliability and validity. *Educational and Psychological Measurement*, 1968, *29*, 665-680.

Sax, G., & Cromack, T. R. The effects of various forms of item arrangements on test performance. *Journal of Educational Measurement*, 1966, *3*, 309-311.

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 1955, *19*, 321-325.

Shavelson, R. J. Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology*, 1972, *63*, 225-234.

Shoemaker, D. M. Allocation of items and examinees in estimating a norm distribution by item-sampling. *Journal of Educational Measurement*, 1970, *7*, 123-128.

Shoemaker, D. M. Criterion-referenced measurement revisited. *Educational Technology*, 1971, *9*, 61-62.

Shoemaker, D. M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. *Journal of Educational Measurement*, 1971, *8*, 215-220.

Shoemaker, D. M. Improving criterion-referenced measurement. *Journal of Special Education*, 1972, *6*, 315-323.

Shoemaker, D. M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. *Journal of Educational Measurement*, 1970, *7*, 255-262.

Shoemaker, D. M., & Osburn, H. G. An empirical study of generalizability coefficients for unmatched data. *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 239-246.

Shuford, E. H., Albert, A., & Massengill, H. E. Admissable probability measurement procedures. *Psychometrika*, 1966, *31*, 125-146.

Sirotnik, K. An investigation of the context effect in matrix sampling. *Journal of Educational Measurement*, 1970, *7*, 199-207.

Sitgreaves, R. A statistical formulation of the attenuation paradox in test theory. In H. Solomon, *Studies in item analysis and prediction*. Stanford, Calif.: Stanford University Press, 1961.

Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: Unsolved problems in test development, assembly and interpretation. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion referenced measurement*. CSE Monograph Series in Evaluation No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Slakter, M. J. Risk-taking on objective examinations. *American Educational Research Journal*, 1967, *4*, 31-43.

Slakter, M. J. The effect of guessing strategy on objective test scores. *Journal of Educational Measurement*, 1968, *5*, 217-221.

Slakter, M. J. The penalty for not guessing. *Journal of Educational Measurement*, 1968, *5*, 141-144.

Smith, R. An empirical examination of the assumptions underlying the "Taxonomy of educational objectives: Cognitive domain." *Journal of Educational Measurement*, 1968, *5*, 125-128.

Smouse, A. D., & Munz, D. C. The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology*, 1968, *68*, 181-184.

Soderquist, H. O. A new method of weighting scores in a true-false test. *Journal of Educational Research*, 1936, *30*, 290-292.

Stalnaker, J. M. Weighting questions in the essay-type examination. *Journal of Educational Psychology*, 1938, *29*, 481-490.

Stanley, J. C. 'Psychological' correction for chance. *Journal of Experimental Education*, 1954, *22*, 297-298.

Stanley, J. Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Stoker, R. W., & Kropp, R. P. Measurement of cognitive process. *Journal of Educational Measurement*, 1964, *1*, 39-42.

Stokes, R. R. The split-response technique. *Phi Delta Kappan*, 1966, *47*, 271-272.

Storey, A. G. Review of evidence or the case against the true-false item. *Journal of Educational Research*, 1966, *59*, 282-285.

Sukeyori, S. A study of readability measurement—applications of cloze procedure to Japanese language. *Japanese Journal of Psychology*, 1957, *28*, (English abstract).

Swineford, F. Note on "Tests of the same length do have the same standard error of measurement." *Educational and Psychological Measurement*, 1959, *19*, 241-242.

Taylor, W. L. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 1953, *30*, 415-433.

Taylor, W. L. Cloze readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology*, 1957, *41*, 12-26.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), *Educational Measurement*. Washington, D. C.: American Council on Education, 1951.

Thorndike, R. L. The problem of guessing. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Torgerson, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

Traub, R. E., & Hambleton, R. K. The effects of scoring instructions and degree of speededness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement*, 1972, *32*, 737-758.

Tryon, R. C. Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin*, 1957, *54*, 229-249.

Tucker, L. R. A note on the estimation of test reliability by the Kuder-Richardson Formula (20). *Psychometrika*, 1949, *14*, 117-119.

Tucker, L. R. Maximum validity of a test with equivalent items. *Psychometrika*, 1946, *11*, 1-13.

Tucker, L. Scales and minimizing the importance of reference groups. In *Proceedings, invitational conference on testing problems.* Princeton, N. J.: Educational Testing Service, 1952.

Tversky, A. On the optimal number of alternatives of a choice point. *Journal of Mathematical Psychology*, 1964, *1*, 386-391.

Votaw, D. F. The effect of do-not-guess directions upon the validity of true-false or multiple choice tests. *Journal of Educational Psychology*, 1936, *28*, 698-703.

Wang, M. C., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, *40*, 663-705.

Webster, H. A generalization of Kuder-Richardson reliability formula 21. *Educational and Psychological Measurement*, 1960, *20*, 131-138.

Webster, H. Item selection methods for increasing test homogeniety. *Psychometrika*, 1957, *22*, 395-403.

Webster, H. Maximizing test validity by item selection. *Psychometrika*, 1956, *21*, 153-164.

Wedman, I. Reliability, validity and discrimination measures for criterion-referenced tests. *Educational Reports*, Umea, No. 4, 1973.

Werts, C. E., & Linn, R. L. Corrections for attenuation. *Educational and Psychological Measurement*, 1972, *32*, 117-127.

Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D. C.: American Council on Education, 1971.

West, P. The significance of weighted scores. *Journal of Educational Psychology*, 1924, *15*, , 302-308.

White, B. W., & Saltz, E. The measurement of reproducibility. *Psychological Bulletin*, 1957, *54*, 81-99.

Whitney, D. R. S and $\chi^2$ tests of association: An empirical comparison. *American Educational Research Journal*, 1972, *9*, 113-122.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 1938, *3*, 23-40.

Wood, R. Response-contingent testing. *Review of Educational Research*, 1973, *43*, 529-544.

Wood, R. The efficacy of tailored testing. *Educational Research*, 1969, *11*, 219-222.

Woodbury, M. A., & Novick, M. R. Maximizing the validity of a test battery as a function of relative test lengths for a fixed total testing time. *Journal of Mathematical Psychology*, 1968, *5*, 242-259.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 63-64.

Woodson, M. I. C. E. The issue of item and test variance for criterion-referenced tests: A reply. *Journal of Educational Measurement*, 1974, *11*, 139-140.

Wright, B., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.